Erik BURY

KU LEUVEN

ARENBERG DOCTORAL SCHOOL
FACULTY OF ENGINEERING SCIENCE

ASSESSING BIAS TEMPERATURE INSTABILITIES AND SELF-HEATING EFFECTS IN ADVANCED SEMICONDUCTOR NODES

October 2016

# Assessing Bias Temperature Instabilities and Self-Heating Effects in Advanced Semiconductor nodes

**Erik BURY**

Supervisor:
Prof.dr. Guido Groeseneken

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Engineering Science: Electrical Engineering
October 2016

# ASSESSING BIAS TEMPERATURE INSTABILITIES AND SELF-HEATING EFFECTS IN ADVANCED SEMICONDUCTOR NODES

Erik BURY

Supervisor:
Prof. dr. G. Groeseneken

Members of the
Examination Committee:
Prof. dr. M. Heyns
Prof. dr. G. Gielen
Prof. dr. G. Ghibaudo
Dr. B. Kaczer
Dr. M. Badaroglu

Dissertation presented
in partial fulfilment of the
requirements for the
degree of Doctor of Doctor
of Engineering Science:
Electrical Engineering

October 2016

© 2016 KU Leuven, Science, Engineering & Technology
Uitgegeven in eigen beheer, Erik Bury, Leuven, België.

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaandelijke schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

# Acknowledgements

Writing the acknowledgements is the last *Chapter* of my PhD. While writing this text, I realized how many people have helped me throughout this research and supported me in my life in general. I owe all of them, each in their way, a debt of gratitude.

I am grateful to Prof. Guido Groeseneken for giving me the opportunity to start this PhD at imec, and for being ever so kind to show interest in my work and for giving his precious and kind advice regarding the topic of my research.

I owe a deep sense of gratitude to Dr. Ben Kaczer, who has been my daily supervisor and who's expertise, understanding and grateful guidance made it possible to me to research topic that are of great interest for me. He provided me with material, information and ideas that I could not possibly have discovered on my own.

I am extremely thankful to Dr. Robin Degraeve, Dr. Jacopo Franco and Philippe Roussel, especially for them being a source of motivation, their endless devotion and many vivid discussions we had on all manner of things. Thanks to Dr. Dimitri Linten for his willingness to proof read countless pages of this thesis.

Many thanks to Prof. Dragica Vasileska and Prof. Katerina Raleva, for the fruitful discussions and the insight they gave me in their work.

Then there is ~~Pieter~~, Dr. Pieter Weckx, a true friend and colleague who started his PhD together with me. I have had the honor to be his *best man* on his marriage, to see him graduate and to see him getting his first child, Mila. Throughout the past years, he never failed to amaze me with his great sense of humor but also his perseverance.

Completing this work would have been all the more difficult were it not for the support and friendship of my colleagues: Simon, Vamsi, Marko, Kent, Adrian, Alexandre, Geert, Roman, Mirko, Shih-Hung, Bart, Doyoung,… and many more, for their great sense of humor and creating a great work environment. I'd also like to mention my former colleagues: Maria Toledano-Luque, Moonju Cho, Thomas Kauerauf and Marc Aoulaiche who were magnificent guides throughout the first years of my PhD.

I would like to thank my brother Bart, my sisters Veerle, Katleen and Klara, and most of all my parents Erna and Hans for their ~~interest in my work~~ unconditional love and support throughout the past 27 years of my life.

Finally, I must express my gratitude to my beloved Charlotte, for her continued support and encouragement. I was continually amazed by her patience, as she experienced all of the ups and downs of my research. Meanwhile, she gave me a fantastic daughter: Lauranne. I dedicate this work to both.

# Abstract

In order to meet the specifications in terms of drive current and electrostatic channel control of nanoscale metal-oxide-semiconductor field-effect transistors (MOSFET), scaling of the equivalent oxide thickness (EOT) is essential. However, with EOT scaling down to dimensions of only a few atomic layers, the reliability of these dielectrics start to become an issue. One of the main MOSFET degradation phenomena is Bias-Temperature-Instability (BTI), which has evolved in a way that the industry's reliability targets can no longer be met with planar devices. In order to maintain electrostatic control without scaling EOT, recently 3D device architectures such as FinFETs and gate-all-around nanowires (GAA NW) were introduced. These geometric modifications raised concern over device self-heating effects (SHE). Moreover, in future technologies, completely new channel materials with high carrier mobility will be utilized.

In this Thesis, we study the BTI reliability by developing a new technique that allows us to quickly screen the effect of tuned process parameters on the BTI resilience of sub-nm EOT dielectrics. Using the gathered information from systematic benchmarking, we have strong indications that the oxide scavenging technique used to form these UT-EOT devices, also inherently forms a fundamental obstacle for BTI reliability because of the defects generated during this processing. We find that the defectivity and the BTI reliability can be improved by modifying the annealing techniques.

Subsequently, more fundamental understanding of BTI-induced oxide defects is provided and by studying its co-interaction with failure mechanisms such as Random-Telegraph-Noise (RTN) and Stress-Induced-Leakage Current (SILC). We demonstrate how gate leakage and fluctuations and charge trapping are related and show that the multi-state non-radiative multi-phonon (NMP) theory can be applied to explain the defect properties.

Thereafter, we develop a measurement methodology to quantify the SHE in planar, FinFET and GAA-NW FETs. The technique is corroborated with electro-thermal simulations, uncovering the asymmetric heating of the device. We propose simulations using thermal conductivity tensors to superseed the underlying phonon scattering physics and also review the impact in circuits.

Finally, concerning future device nodes with high mobility channels, we find that introducing non-dilute alloys, for example to enhance strain in pFET devices, has a strong impact on the SHE.

# Beknopte samenvatting

Om de vooropgestelde specificaties van stroom en elektrostatische kanaal-controle van metaal-oxide-halfgeleider veld-effect transistoren op nanoschaal te kunnen behouden, is het schalen van de equivalente oxide dikte (EOT) essentieel. Echter, wanneer deze EOT de dimensies van slechts enkele atomaire lagen bereikt, vormt de betrouwbaarheid van deze diëlektrica een probleem. Eén van de belangrijkste degradatiefenomenen is instelpunt-temperatuur-instabiliteit (Engels: Bias Temperature Instability), dewelke intussen geëvolueerd is op het punt dat de betrouwbaarheidscriteria die gespecifieerd werden door de industrie niet langer gehandhaafd kunnen blijven in vlakke transistoren. Om de elektrostatische controle te kunnen behouden zonder het schalen van de EOT, heeft de industrie geopteerd om over te gaan naar nieuwe transistorarchitecturen zoals vinvormige (Engels: FinFET) en ronde-poort (Engels: Gate-All-Around) transistoren. Deze wijzigingen aan de geometrie hebben de bezorgdheid over zelf-geïnduceerde hitte-effecten (Engels: Self-Heating Effects) doen toenemen. Bovendien zal in toekomstige transistoren gebruik gemaakt worden van nieuwe kanaalmaterialen met hoge-mobiliteitseigenschappen.

In deze Thesis bestuderen we de BTI-betrouwbaarheid door het ontwikkelen van een nieuwe meettechniek die ons toelaat om snel de effecten van aangepaste proces-parameters op de BTI-resistentie van sub-nanometer EOT diëlektrica te evalueren. Op basis van de systematisch verzamelde meetinformatie, hebben we sterke indicaties dat de oxide-opruimingstechniek (Engels: scavenging), die gebruikt wordt om deze ultra-dunne EOT MOSFETs te vormen, ook inherent een fundamenteel obstakel vormt voor de BTI-betrouwbaarheid omwille van de defecten die hierdoor gegenereerd worden tijdens de fabricatie. We vinden echter wel dat de defectiviteit kan verminderd worden door het aanpassen van de tempercondities.

Vervolgens wordt een meer fundamenteel begrip van de BTI-geïnduceerde oxide defecten gepresenteerd door het bestuderen van de co-interactie met andere faalmechanismen zoals ruis (Engels: Random-Telegraph-Noise) en stress-geïnduceerde lekstroom (Engels: Stress-Induced-Leakage Current). We tonen aan hoe poort lekstromen en –fluctuaties, en ladingsvangst gecorreleerd zijn en tonen aan dat de theorie betreffende de meervoudige-staat van niet-radiatieve multi-fononen emissie (Engels: multi-state non-radiative

multiphonon emission) toegepast kan worden om de eigenschappen van de defecten te beschrijven.

Nadien ontwikkelen we een meetmethodologie om de zelf-geïnduceerde hitte-effecten te kwantificeren in vlakke, Fin- en GAA FETs. Deze techniek wordt bekrachtigd met elektro-thermische simulaties, dewelke de asymmetrische warmtegeneratie van de transistor blootlegt. We stellen simulaties voor die gebruikmaken van temperatuursafhankelijke tensoren om de onderliggende fysica van fononentransport en –botsingsmechanismen na te bootsen en bestuderen tevens de gevolgen van SHE op circuitniveau.

Tot slot, betreffende het gebruik van hoge-mobiliteitskanalen in toekomstige technologieën, stellen we vast dat bij het gebruikmaken van niet-verdunde legeringen, zoals bijvoorbeeld Silicium-Germanium om een hogere mechanische spanning te bekomen in pFET transistoren, een sterk effect heeft op de zelf-geïnduceerde hitte-effecten.

# List of Symbols

| Symbol | Unit | Meaning |
|---|---|---|
| $C_d$ | $F/cm^2$ | Depletion layer capacitance |
| $C_{GA}$ | $F/cm^2$ | Gate-to-all capacitance |
| $C_{GB}$ | $F/cm^2$ | Gate-to-bulk capacitance |
| $C_{GC}$ | $F/cm^2$ | Gate-to-channel capacitance |
| $C_K$ | - | Empirical alloying bowing factor |
| $C_{OX}$ | $F/cm^2$ | Oxide capacitance |
| $C_s$ | $J/m^3K$ | Volumetric heat capacity |
| $C_{TH}$ | $J/K$ | Thermal capacitance |
| $D_{IT}$ | $cm^{-2}$ | Interface trap density |
| $D_N$ | - | Diffusion constant of carriers |
| $E$ | $V/cm$ | Lateral electric field |
| $E$ | $eV$ | Carrier energy |
| $E_A$ | - | Arrhenius activation energy |
| $E_G$ | $eV$ | bandgap |
| $f$ | $Hz$ | Frequency |
| $F$ | $V/cm$ | Electric field |
| $F_M$ | $V/cm$ | Maximum electric field in the channel |
| $g_{ds}$ | $S$ | Output conductance |
| $I$ | $A$ | Current |
| $I_D$ | $A$ | Drain current |
| $i_{dis}$ | $A$ | Displacement current |
| $I_{DSAT}$ | $A$ | Saturation drive current |
| $I_{SUB}$ | $A$ | Substrate current |
| $k$ | $F/m$ | Dielectric permittivity |
| $k_B$ | $eV/K$ | Boltzmann constant |
| $k_{TH}$ | $W/mK$ | Thermal conductivity |
| $L_G$ | $nm$ | Gate length |
| $M$ | - | Multiplication factor |
| $n$ | - | Time acceleration exponent |
| $N$ | $C/cm^2$ | Sheet charge density |
| $N_A$ | $cm^{-3}$ | Acceptor dopant concentration |
| $N_D$ | $cm^{-3}$ | Donor dopant concentration |
| $N_{FIN}$ | - | Number of fins |
| $n_i$ | $cm^{-3}$ | Intrinsic Si carrier concentration |
| $N_{IT}$ | - | Number of interface states |

| | | |
|---|---|---|
| $N_T$ | - | Number of traps |
| $P$ | - | Permanent component of threshold voltage shift |
| $P$ | $W$ | Power |
| $P_N$ | - | Probability for having N traps in a device |
| $q$ | $C$ | Elementary electron charge |
| $q$ | $W/m^2$ | Heat flux density |
| $Q$ | $W$ | Total heat flux |
| $Q_{IT}$ | $C$ | Interface trapped charge |
| $Q_{OT}$ | $C$ | Oxide trapped charge |
| $Q_{trapped}$ | $C$ | Trapped charge |
| $R$ | - | Recoverable component of threshold voltage shift |
| $R_{TH}$ | $W/mK$ | Thermal resistance |
| $T$ | $K$ | Temperature |
| $T_0$ | $K$ | Room temperature |
| $T_{inv}$ | $nm$ | Inversion layer thickness |
| $T_{ox}$ | $nm$ | Gate dielectric thickness |
| $t_r$ | $s$ | Relaxation time |
| $t_s$ | $s$ | Stress time |
| $t_{stress}$ | $s$ | Cumulated stress time |
| $V_{BD}$ | $V$ | Breakdown voltage |
| $V_{DD}$ | $V$ | Circuit operating voltage |
| $V_{FB}$ | $V$ | Flatband voltage |
| $v_{inj}$ | $cm/s$ | Source injection velocity |
| $V_{OV}$ | - | Overdrive voltage |
| $V_{TH}$ | $V$ | Threshold voltage |
| $W$ | $nm$ | Device width |
| $W_{dep}$ | $nm$ | Depletion region thickness |
| $W_{FIN}$ | $nm$ | Fin width |
| $\alpha$ | - | Scale parameter |
| $\alpha_{vsat}$ | $m/sK$ | Saturation velocity temperature coefficient |
| $\alpha_{VTH}$ | $m/VK$ | Threshold voltage temperature coefficient |
| $\beta$ | - | Weibull-modulus |
| $\gamma$ | - | Voltage acceleration exponent |
| $\varepsilon_0$ | $F/m$ | Absolute permittivity |
| $\varepsilon_r$ | - | Relative dielectric permittivity |

| | | |
|---|---|---|
| $\zeta_{defect}$ | $eV$ | defect band offset w.r.t. conduction band |
| $\eta$ | $V$ | Mean impact per defect |
| $\kappa$ | - | Voltage scale factor |
| $\kappa$ | - | Thermal conductivity |
| $\lambda$ | - | Linear dimension scale factor |
| $\lambda$ | $nm$ | Carrier mean-free-path |
| $\Lambda$ | $nm$ | Average phonon mean free path |
| $\mu$ | $cm^2/Vs$ | Carrier mobility |
| $\nu$ | $cm/s$ | Phonon velocity |
| $\xi$ | - | Universal relaxation time |
| $\sigma$ | $S$ | Electrical conductivity |
| $\tau$ | $s$ | Total device lifetime |
| $\tau_c$ | $s$ | Capture time |
| $\tau_{channel}$ | $s$ | Channel heating time constant |
| $\tau_e$ | $s$ | Emission time |
| $\phi_b$ | $eV$ | Barrier for de-passivating a Si-H bond |
| $\phi_f$ | $eV$ | Fermi level energy |
| $\phi_i$ | $eV$ | Energy threshold |
| $\phi_{it}$ | $eV$ | Energy threshold to create an interface trap |
| $\phi_t$ | $mV$ | Thermal voltage |
| $\chi$ | $eV$ | Band offset w.r.t. conduction band |
| $\Psi$ | $eV$ | Band bending |

# List of Acronyms

| | |
|---|---|
| ALD | Atomic-layer-deposition |
| BEOL | Back-end-of-line |
| BTI | Bias temperature instability |
| CBCM | Charge-based capacitance measurement |
| CC | Cold-carrier |
| CESL | Contact-etch-stop-layer |
| CHC | Channel-hot-carrier |
| CHE | Channel hot-electron |
| CMOS | Complementary Metal-oxide-semiconductor |
| CNT | Carbon nanotube |
| C-V | Capacitance-Voltage |
| DAHC | Drain avalanche hot-carrier |
| DT | Direct tunneling |
| DUT | Device-Under-Test |
| EKV | Enz, Krummenacher and Vittoz |
| eMSM | extended Measure-Stress-Measure |
| EOT | Equivalent oxide thickness |
| EPI | Epitaxial |
| eSiGe | embedded SiGe |
| ET | Electro-thermal |
| EWF | Electric work function |
| FD | Fully-depleted |
| FEOL | Front-end-of-line |
| FET | Field-effect-transistor |
| FN | Fowler-Nordheim |
| GAA | Gate-all-around |
| GF | Gate first |
| GL | Gate last |
| HBD | Hard breakdown |
| HF-CV | High-frequency C-V |
| HK/MG | High-k/Metal Gate |
| HKF | High-k first |
| HKL | High-k last |
| HRTEM | High-resolution TEM |
| IC | Integrated Circuit |
| IL | Interfacial layer |

| | |
|---|---|
| ILD | Interfacial layer deposition |
| ISSG | In-situ steam generation |
| ITRS | International Technology Roadmap for Semiconductors |
| *I-V* | Current-Voltage |
| LEM | Lucky Electron Model |
| LER | Line-edge-roughness |
| LRME | Lattice relaxation multi-phonon emission |
| MC | Monte Carlo |
| MGG | Metal-gate-granularity |
| MIPS | Metal-inserted poly-Si |
| MOS | Metal-oxide-semiconductor |
| MOSFET | Metal-oxide-semiconductor field-effect-transistor |
| MuGFET | Multi-gate field-effect-transistor |
| MVE | Multi-vibrational modes |
| NBTI | Negative bias temperature instability |
| NMP | Non-radiative multi-phonon |
| NW | Nanowires |
| PBTE | Phonon Boltzmann transport equation |
| PBTI | Positive bias temperature instabilities |
| PIV | Pulsed-IV |
| PMA | Post-metallization anneal |
| PTM | Predictive technology model |
| PVD | Physical Vapor deposition |
| QW | Quantum well |
| R-D | Reaction diffusion |
| RF | Radio-frequency |
| RMG | Replacement gate |
| ROCP | Ring-oscillator charge pumping |
| RTA | Rapid-thermal-anneal |
| RTO | Rapid-thermal oxidation |
| SGHE | Secondary generated hot-electron |
| SHE | Self-heating effect |
| SHE | Substrate hot-electron |
| SILC | Stress-induced-leakage current |
| SMU | Source Measurement Unit |
| SOI | Silicon-on-Isolator |

| | |
|---|---|
| SRAM | Static random-access memory |
| SRB | Strain-relaxed buffer |
| SS | Subthreshold swing |
| TAT | Trap-assisted tunneling |
| TCR | Temperature coefficient of resistance |
| TDDB | Time-dependent-dielectric-breakdown |
| TDDS | Time-dependent-defect-spectroscopy |
| TEM | Transmission electron microscopy |
| TFET | Tunnel FET |
| TSCIS | Trap-spectroscopy by charge injection and sensing |
| TSV | Through-Silicon-Via |
| ULSI | Ultra-large-scale-integration |
| UTBB | Ultra-thin body and box |
| UT-EOT | Ultra-Thin Equivalent-Oxide-Thickness |
| VLS | Vapor-liquid-solid |
| VNA | Vectorial network analyzer |
| WF | Work function |

# Table of Contents

# Chapter 1: Introduction

The technological developments in the semiconductor industry enabled the wide-spread use of electronics and computer-aided systems in almost all aspects of our lives, being it in corporate, household or bio-medical applications,... we can safely state that micro-electronics have become ubiquitous.

The cornerstone of the above described developments relies on the invention and elaboration of the MOS transistor in the last decades. For this reason, this chapter will provide the reader with an overview of the major advancements in semiconductor device development and scaling ever since.

Together with these advancements however, semiconductor device designers have always been concerned with device reliability issues. One of the hot topics in the silicon industry is the issue of growing or depositing a stable and reliable gate dielectric. And even though these reliability issues never stopped the continuing downscaling, it unmistakably steered evolution of the MOSFET.

For that reason, the reader will also be given an insight in *upcoming* innovations in semiconductor device development for the future generations of technology ("nodes") and we will address some of the major challenges that are expected to come along with aggressive scaling of silicon devices.

## 1.1 The discovery of the MOS transistor

It is almost 70 years ago that Shockley, Bardeen and Brattain (Fig. 1) created the first working bipolar point junction transistor in the Bell Laboratories [Riordan99]. It is this transistor that we can call the "*nerve cell*" of the Information Age. The transistor was able to amplify and switch electronic signals and electrical power.

Remarkably, already the prototypes of this transistor, suffered from reliability issues, very similar to those that are still seen nowadays: the lack of significant current modulation was attributed to "*surface states*". It was only after Brattain found out that immersing a silicon-semiconductor in an electrolyte could neutralize the effect of these field-blocking "*surface states*", that the first working transistor could be produced [Compu15].

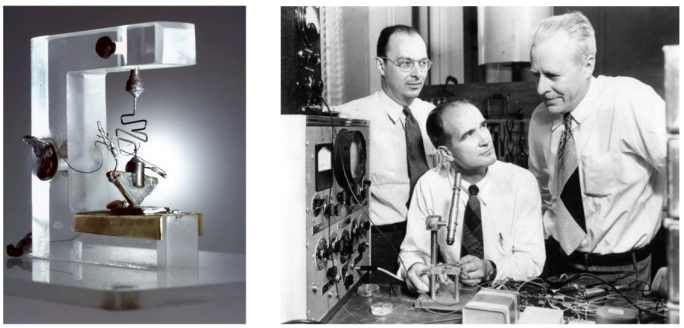


**Fig. 1: Replica of the first point-contact transistor, as invented by (from left to right) Bardeen, Shockley and Brattain in 1947. [copied from: http://www.ecse.rpi.edu/~schubert ]**

In the subsequent period, many new transistor structures were studied. Shortly after, Shockley invented and patented the junction-transistor, the prototype of the bipolar transistor [Shockley51]. During the 1950s, these junction semiconductor devices gradually replaced vacuum tubes in digital computers.

The invention of the transistor was however not sufficient to start the evolutionary era of transistor scaling as we know it. The *tyranny of numbers* was a problem faced in the 1960s by computer engineers: engineers were unable to increase the performance of their designs due to the huge number of components involved [Ti15]. Every component needed to be wired to many other components. The latter were typically soldered together by hand. If the scientists at that time wanted to improve further performance, more components would be needed, and it seemed that future designs would consist almost entirely of wiring.

It was in the late 50's that one came up with the idea that the semiconductor material itself—germanium at that time—could be used to make all common electronic components on the same base (or substrate material) and interconnected without wires. Even though there is no consensus, Jack Kilby at Texas Instruments and Robert Noyce at Fairchild are mostly credited for having invented the first integrated circuit (IC) in 1958 and 1959.

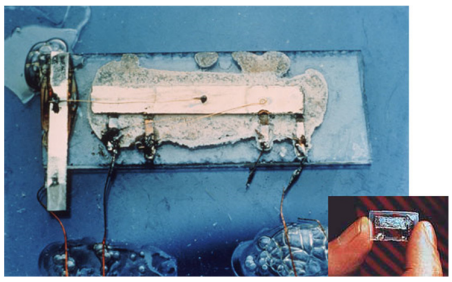


**Fig. 2: The original integrated circuit of Jack Kilby [copied from http://history-computer.com/ModernComputer/Basis/IC.html ]**

The first prototype IC constructed by Kilby in 1958 contained only one transistor, several resistors, and a capacitor on a single slab of germanium, and featured fine gold wires to interconnect them (Fig. 2). However, because the wires still had to be individually attached, this type of design was not practical to manufacture. Noyce on the other hand, patented his "planar" IC design in 1959, where all the components are diffused in or etched on a silicon base, including a layer of aluminum metal interconnects. And so, in 1960, Fairchild (the company where Noyce was working) constructed the *first planar IC*, consisting of a flip-flop circuit with four transistors and five resistors on a circular substrate of about 20 $mm^2$ in size [Moore98].

It was also in the 60's that Atalla and Kahng, also at Bell Labs, achieved the first successful attempt of an insulated-gate field-effect-transistor (FET). Even though the working principle of the FET had been described before [Lilienfeld28], the device appeared to be even more susceptible to "*surface states*". Investigating thermally grown silicon-dioxide layers and special cleaning procedures, they found these states were markedly reduced by the effect of surface passivation at the interface between the silicon and its oxide in a sandwich comprising layers of **M**etal, **O**xide, and **S**ilicon - thus the name MOSFET, now popularly known as MOS [Arns98].

It became quickly clear that the MOSFET could become the number one alternative for the high power-consuming, bulky and expensive vacuum-tubes, used in computers at that time.

## 1.2 Smaller, faster, cheaper

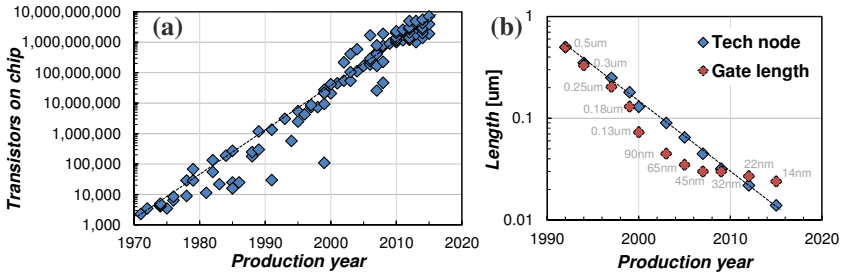



**Fig. 3: (a) 30 years after Moore's famous publication, Moore's law still appears to be valid but (b) not all transistor dimensions, such as the gate length, are scaling accordingly anymore due to physical scaling limits. [(a) replotted from Wikipedia-article "transistor count" and (b) partially replotted from: Thompson02, Ghani03, Chau04, Bai04, Mistry07, Jan09, Auth12, and eetimes14]**

Discussed in Section 1.1, the use of several hundreds of individually fabricated transistors, capacitors, resistances and other elements was labor-intensive in assembly and insufficiently reliable.

The true success of the semiconductor industry relies on the continuous performance improvements of the integrated circuits. This is mainly achieved by reducing the dimensions of the key component: the MOSFET. Reducing the device dimensions allowed the integration of a larger number of transistors on a chip as well higher speeds and increased functionality.

The progress of the microelectronics industry, was also predicted by Moore, one of the cofounders of Intel. Based on early trends, he predicted that the transistor density that could be produced on an IC would obey an exponential increase as a function of time [Moore65], popularly known as Moore's law. At the time of the publishing of this work, it has been exactly 50 years since then and surprisingly, this law has been followed over almost perfectly by the microelectronics industry. However, in part it is also a self-fulfilling prophecy because this law has been depicting the long-term goals of the semiconductor industry and it sets the targets for research and development.

**Table I: Dennard's scaling theory to MOSFET parameters for constant voltage and constant field scaling. Note that the effects of velocity saturation are not taken into account. [replotted after Dennard74]**

| | Voltage sc. | Field sc. |
|---|---|---|
| ***Dimension:*** | $\lambda\kappa$ | $\kappa$ |
| ***Potential:*** | $\kappa$ | $\kappa$ |
| ***Impurity concentration:*** | $\kappa/\lambda^2$ | $1/\kappa$ |
| ***Electric Field:*** | $\kappa/\lambda$ | **1** |
| ***Oxide capacitance:*** | $\lambda$ | $\kappa$ |
| ***Current:*** | $\kappa^2/\lambda$ | $\kappa$ |
| ***Power:*** | $\kappa^3/\lambda$ | $\kappa^2$ |
| ***Power * delay:*** | $\lambda^2/\kappa$ | $\kappa$ |

**$\kappa$ = voltage scale factor**

**$\lambda$ = lineair dimension scale factor**

Concretely, Moore's law predicts that in every 18 month cycle[1], the transistor density would double, thereby requiring transistor dimensions to shrink approximately 30% for every technology node (Fig. 4(a)). Starting in the 70's from Intel's 4004 micro-processor, scaling continued ever since. Investigating the supply voltage scaling trend can yield much insight in the MOSFET scaling past, present and future.

The initial scaling trend was one of constant supply voltage, mainly to enhance the FET performance as predicted by Dennard's scaling theory, depicted in Table I [Dennard74]. For continuously shrinking horizontal and lateral device dimensions, this signifies a steady increase in electric fields. It became clear that this trend was not sustainable on the long term.

Therefore, this period of constant voltage scaling was followed by an era of constant field scaling in the 90's, in which the supply voltages were scaled according to the shrinking device dimensions. Due to the inability of reducing the threshold voltage of the devices, caused by the non-scaling sub-threshold slope of the MOSFET, the supply voltages have more or less saturated to scale

[1] In his 1965 paper, Moore predicted a 12 month cycle, but he adjusted his prediction to 18 months in 1970.

since the 65nm node around 1V [Groes10]. As a result, the electric fields were increasing again in ranges up to 10MV/cm and 1MV/cm for the oxide fields and lateral fields respectively, putting engineers in to trouble maintaining the reliability of the devices. These trends are depicted in Fig. 4.

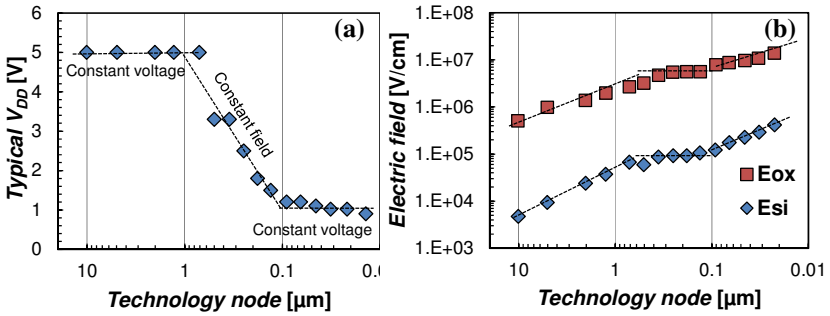



**Fig. 4: (a) Evolution of operating voltages of various technology nodes over the last decades shows clearly three periods in the era of scaling. (b) The resulting electric fields in the device are increasing steadily since the 0.130µm node [replotted from Franco11 and Groeseneken10].**

However, noticing that the gate length was about 30nm for the 32nm node, this would mean that in a few generations from then the gate length would scale down to the atomic level, which is a physical limit for scaling. And indeed, in the most recent nodes, we observe that the gate length is no longer scaling accordingly (Fig. 3(b). Since the 90nm node, we can state that the gate length only scaled according to ~0.9 per node. However, by scaling the gate pitch, the overall area scaling of ~0.5 per node could be maintained, as predicted by Moore's law.

One of the other important scaling factors is the gate dielectric thickness (Tox). Thinner gate dielectrics allow greater capacitive control over the transistor channel [Kuhn09]. This results in turn in higher drive currents and thus higher performance. It is this gate oxide thickness reduction that caused the oxide field to increase steadily.

The reduction of the gate oxide thickness is also one of the main reasons that caused the overall power density of the devices to increase continuously: the gate leakage increases *exponentially* by about 1 order of magnitude for every 2Å of gate oxide thickness reduction in the direct tunneling regime (i.e.

for thicknesses below 5 nm). As a result, power densities of more than 100W/cm² were already reached in the early 2000's, which caused issues of readily overheating chips, as illustrated in Fig. 5. This issue is now mostly tackled with the integration of on-chip temperature sensors and a dynamic control of operating voltages and frequencies.

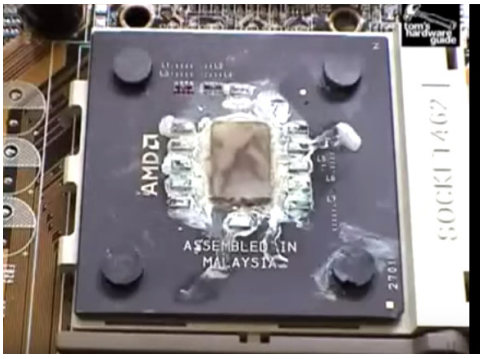



**Fig. 5: The high power dissipation of CPU's in the early 2000's could cause chips, lacking build-in temperature diodes at that time, to burn within seconds after the cooling heatsink was removed. [source: tomshardware.com]**

Continuing the scaling process, it became clear that the gate leakage would even become dominant over the subthreshold leakage if the gate oxide continued to scale at the same pace [Nowak02]. Moreover, the *effective* channel length defined as distance between source and drain junctions, scaled at an even lower pace than the gate 'drawn' length depicted in Fig. 6 due to the shrinking source and drain to gate overlaps.

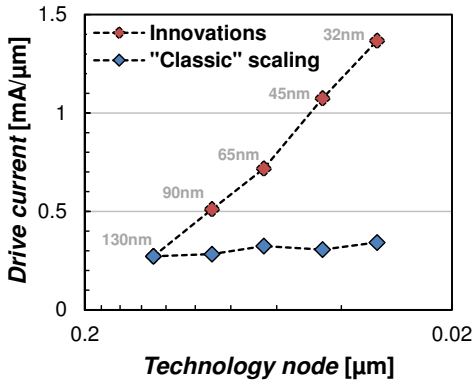



**Fig. 6: Comparison on the evolution of the transistor drive current caused by 'classical' *effective* gate length scaling versus the obtained drive current due to engineering innovations. [replotted from Kuhn09]**

In Fig. 6, the evolution of the transistor drive current with scaling is depicted, if the drive current was only increased due to gate length scaling. This picture makes once again clear that several other significant innovations were needed to prolong the transistor drive current increase, i.e. the golden era of 'Dennard' scaling ended at the 130nm node.

The Section 1.3 discusses the major engineering solutions beyond geometrical scaling that enabled performance enhancements to continue beyond the 130nm node.

## 1.3 Recent transformations of the MOS transistor

From the 130nm node onwards, further scaling of the nodes could only be continued due to the introduction of a few major innovations: the introduction of channel strain for improved carrier mobility, the introduction of high-k gate dielectrics to reduce the gate leakage and finally the introduction of the multi-gate device architecture, allowing improved electrostatic control. As illustrated in Fig. 7, each of these innovations allowed the continuation of the scaling for two or three more technology nodes. It is however clear that these modifications in device architecture are become increasingly radical.

An overview of these innovations is given in the following Section.

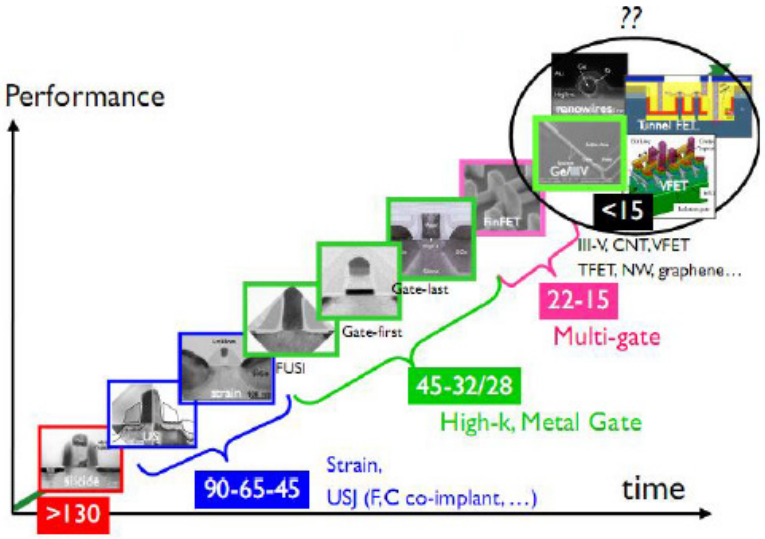



**Fig. 7: Innovations in device engineering that allowed continuous scaling over the last decade. Note the increasing complexity of the engineering solutions. For technology nodes beyond 7nm, it might become insuperable to use devices with completely aberrant operating physics such as tunnel FETs (TFETs).**

### *1.3.1 The 90-65nm node: introduction of stressors*

With the introduction of the 90nm technology node in 2003, the gate length was reduced by 40%, down to 45nm [Ghani03]. As channel control is lost for shorter gate lengths, this had to be compensated with a gate oxide thickness reduction and increased channel doping [Baklan07]. However, the most tolerable gate oxide reduction was only 20% and the additional channel doping would cause the carrier mobility to decrease. The impact of both effects is shown in Fig. 8.

The main reason for the strong gate length reduction compared to the 130nm node, was the application of channel stressors, which were shown only to be effective for strongly scaled gate lengths. In terms of gate oxide thickness, the reduction was hold only by 20%. Still the enhancement in transistor performance could be maintained due to the introduction of channel stressors:

embedded SiGe (eSiGe) source/drain were adopted to generate a highly compressive strain along the channel, due to the lattice spacing of SiGe being larger than Si. For the nMOS devices on the other hand, a SiN Contact Etch Stop Layer (CESL) was developed, where the internal stress within the deposited nitride layer can put tensile strain on the channel [Nara09].

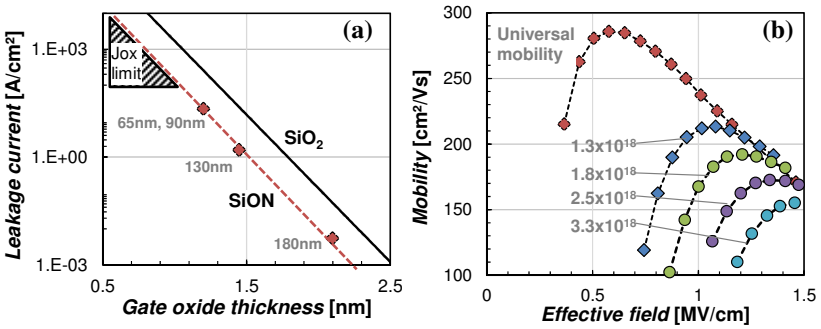



**Fig. 8 (a) Gate leakage current increases about ~10x per 2A oxide thickness reduction. No physical thickness scaling was adopted between the 90nm and 65nm node. (b) The carrier mobility strongly decreases as a function of channel doping due to ionized impurity scattering mechanisms [replotted from Ghani00].**

These stressors showed their effectiveness also for the following-up 65nm node. For the first time, no physical gate oxide scaling was applied, keeping the physical oxide thickness up to 1.2nm. Also the gate length was only reduced by 20%. The reason for the minor gate length scaling, is that if $T_{ox}$ does not scale, scaling the gate length requires higher substrate doping, which degrades both the drive current and the gate-induced-drain-leakage. [Chidam04]. However, the transistor drive currents could still be increased by means of ultra-shallow junctions and enhanced channel strain. TEM micrographs for 130 down to 65nm technology are shown in Fig. 9.

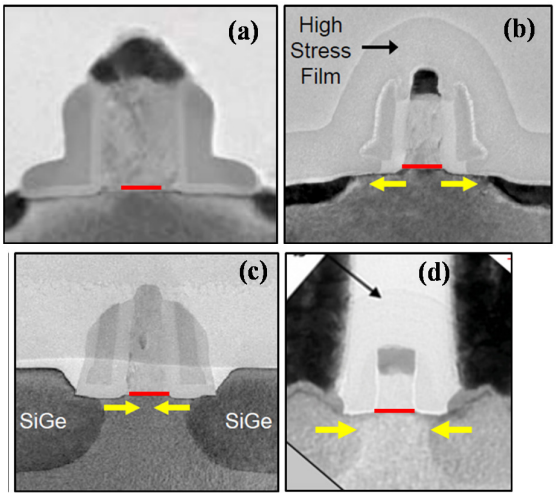



**Fig. 9: TEM micrographs *with identical scales* of Intel's (a) 130nm, (b) 90nm nMOS, strained along the channel with an SiN Contact Etch Stop Layer (CESL), (c) 90nm pMOS, stressed along the channel with SiGe source and drain regions and (d) 65nm pMOS also stressed along the channel. Note the little difference in gate length between the latter two generations. Re-plotted from [Kuhn09] and [Ghani03].**

### *1.3.2 The 45-32nm node: introduction of high-k gates*

As the gate oxide did not scale in the 65nm node, having clear implications for the obtainable drive current and other dimensional scaling of the transistors, solutions had to be found to overcome the $T_{ox}$ scaling limit. The main issue here was the increase in leakage current. Since the 180nm node, SiON, having a 20% higher dielectric constant ($k$) than pure $SiO_2$ worked well enough. A higher dielectric constant helps to achieve the same effective-oxide-thickness (EOT) with a thicker film.

As a solution for the 45nm node, a high-k dielectric material was introduced as an alternative to the SiON, with a 300 to 400% larger $k$ value. In this way, an oxide layer with a thicker physical thickness could be used, yielding the same EOT, thereby strongly reducing the leakage current. For that reason, the

EOT and the physical thickness of the gate stacks are no longer equivalent, as depicted in Fig. 10.

The study of high-k materials was already ongoing for years and a variety of films were investigated before the choices converted towards the Hf-based films, such as $HfO_x$ and $HfSiO_x$ for the 45 and 32 nm nodes. However, these Hf-based films showed to have a much higher defectivity than typical SiON layers. Moreover, depositing the new $HfO_2$ directly on the silicon seemed to be problematic for the interface quality. For that reason, typical gate dielectric films are a hybrid combination of $SiO_2$ near the silicon interface and subsequently $HfO_2$ on top.

A further EOT reduction could be obtained by changing the gate electrode material: in the conventional poly-Si gate a depletion layer is formed which increases the effective thickness. In metal gates however, no depletion layer is formed. For the 45 and 32 nm nodes TiN and TaN were chosen as gate electrodes [Moroz11].

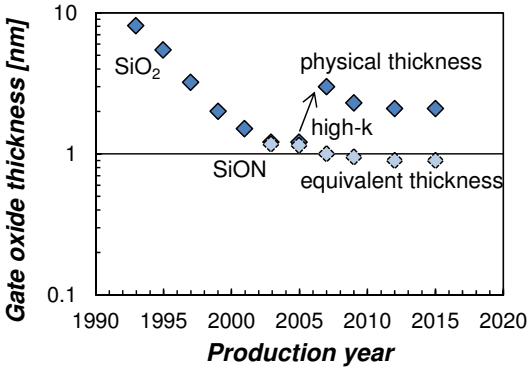



**Fig. 10: Evolution of physical and equivalent oxide thickness over the last two decades of MOSFET scaling shows a clear stagnation after the 90nm node in 2003.**

Along with these new materials, also the process flow that comes along with the production of these devices, did undergo several changes. Nowadays, a typical gate process comes in two flavors: gate-first (GF) and gate-last (GL) [Veloso11]. The gate-first scheme is identical as what has been used in the typical poly-Si processes. In the *gate-first* process, the gate stack has to go

through the high-temperature treatment steps used for dopant activation annealing, silicidation of the contacts, etc.. This causes the work-function to go down which causes an increase of the threshold voltage on its turn. In the *gate-last* process, a dummy gate is processed first and etched away after these high-temperature treatment steps. Only in the last phase, the final gate-stack is processed. This scheme is therefore often referred to as replacement gate (RMG). The latter process has the additional advantage that the dummy-gate removal enhances the strain in the channel.

With the innovation of high-k stacks, the EOT could be reduced to levels which were thought to be fundamental limits for gate leakage current. In those years following, to further increase the gate control for the upcoming nodes, the International Technology Roadmap for Semiconductors (ITRS) predicted that gate dielectrics with an EOT of 0.65 nm and less would be necessary for the future nodes [ITRS]. It was already apparent that the actual scaling of EOT was lagging behind the historical targets. The ITRS re-adjusted their EOT roadmaps year after year (Fig. 11).

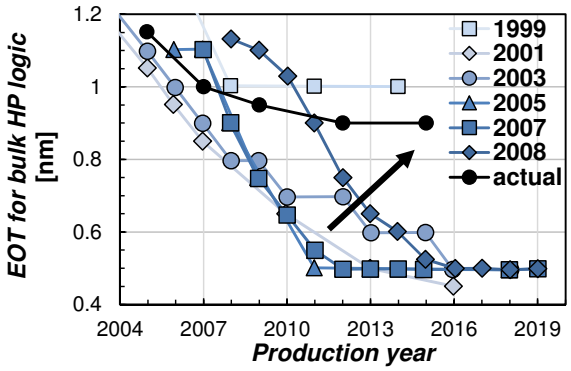



**Fig. 11: Effective-oxide-thickness as predicted by the ITRS roadmap for bulk HP logic over the years shows the slackening and saturation due to the failure of overcoming good solutions. The actual EOT is now even 100% larger than predicted in 2008. [partially replotted from Iwai01].**

The ITRS trends for the EOT of the gate insulator saturated at 0.5 nm. This would cause an increase in the off-leakage current but *decrease* threshold variation in a future small geometry MOSFETs [Toleda11 and Kuhn07]. However, from the actual EOT scaling it is seen that the scaling has been

saturated around 0.9 nm EOT. The main barrier for this are the few atomic layers of SiOx (about 0.2nm/atomic layer) in between the $HfO_2$ and the Si, caused by interfacial and oxide layer defectivity considerations. TEM micrographs for this technology are shown in Fig. 12.

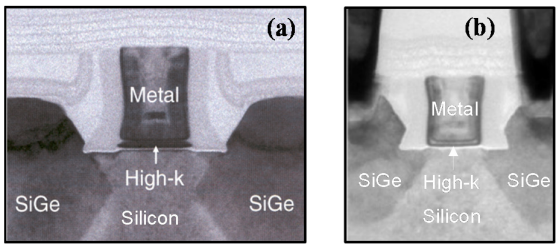



**Fig. 12: TEM micrographs of Intel's (a) 45 and (b) 32nm technology on the same scale, introducing high-k/metal gate dielectrics [replotted from Bohr11].**

From the 32nm node on, gate control became an issue: the current tended to flow deep in the substrate, no longer controlled by the gate. Even though high halo doping was used as a countermeasure, this degraded the on-current and increased the band-to-band-tunneling leakage, as shown in Fig. 13. Band-to-band tunneling happens at the tip of the drain extension, where high field overlaps with high halo doping at the lower corners of the gate-controlled space-charge region.

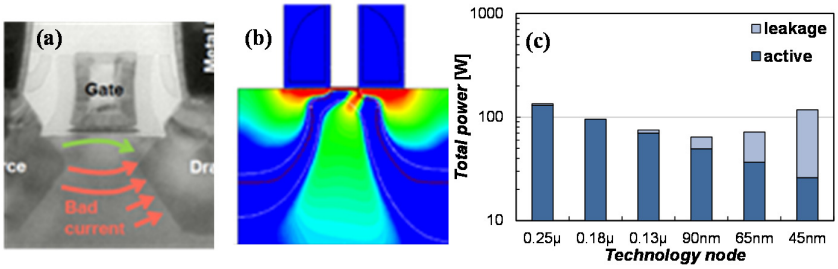



**Fig. 13: (a) Halo doping was used to decrease the drain-induced-barrier-leakage, but this degrades the on-current and (b) increases the band-to-band-tunneling at the drain. (c) Reduced gate control over the channel increases the increases the static leakage power consumption of the processors [replotted from Moroz11, Doyle02].**

### *1.3.3 22nm-14nm node: Introduction of the multi-gate FET*

In order to maintain gate control even though EOT was not scaled, a "multi-gate" (MuGFET) device architecture was introduced by Intel in the 22nm node (Fig. 14 and Fig. 15). Up to that point, the architecture used to be planar, but was replaced by a 3D geometry. As the gate is wrapped around a thin silicon layer, this allows fully-depleted (FD) operation. The regained control of the channel reduces the off-state leakage and allowed further gate length scaling [Colinge97, Park02], albeit moderately (see Fig. 3).

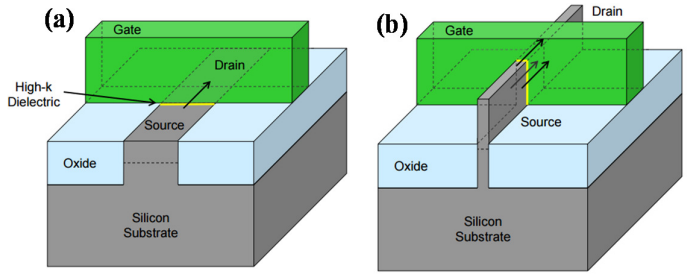



**Fig. 14: Schematic image of (a) a typical planar transistor and (b) Intel's 22nm FinFET [replotted from Auth12]**

FinFETs are however not the only possibility to get a fully-depleted channel. Another well-studied alternative is the use of ultra-thin-body and box (UTBB) transistors. However, even though the manufacturability of these FD-silicon-on-isolator (SOI) devices seems better as the conversion to FinFET requires major changes in process and design, the necessity to use expensive SOI wafers is a major hurdle, even to date.

Apart from the regained electrostatic control, the fully-depleted operation of the transistor allowed the use of a more midgap work function (WF) metal gate, thereby resulting in lower electric fields in the oxide. This allowed Intel to scale the EOT from the 32nm to the 22nm, even though a small relaxation was expected. Finally, as the fin itself is un-doped, it is insensitive to random dopant fluctuations, having positive implications on the transistor $V_{TH}$-variability.

In terms of transistor's performance, the typical metric of 'mA/um' drive current has been adjusted by device engineers. Whereas in the past the current was normalized to the width of the device, initially in FinFET technology, the current was normalized to the circumference of the gate wrapped around the fin. As this metric is slightly subjective (there is no clear boundary until what point the gate is effectively wrapped around the fin), the drive current is now typically normalized by the width of the footprint of the device.

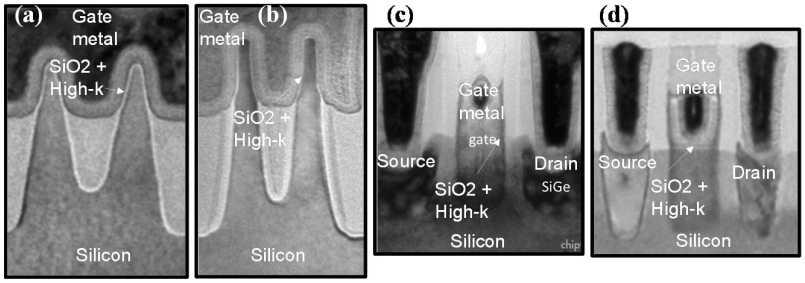



**Fig. 15: TEM along the gate of Intel's (a) 22nm and (b) 14nm FinFETs. Notice the rounded corners to avoid high electric fields at the edges. Intel's second generation FinFETs in (b) are taller. (c) and (d) represent crossections across the gate, similar to previous nodes. Note how the gate length was barely scaled, in contrast to the interconnect density. [replotted from Auth12, Natarajan14 and Semimd12 (22nm TEM)]**

## 1.4 Upcoming innovations in MOS transistors

At the time of writing of this thesis, the 10nm node is expected in 2017. For this node, it is still under debate if we will see another major innovation. Based on the trends that were described in the previous section, we saw a typical two-node cycle for every significant innovation. Also the most recent 14nm node was a logic evolution of its predecessor.

However, as the challenges for process integration are increasing steeply, we might see another Si-based FinFET node first. Indeed, looking at the ITRS 2013 roadmap [ITRS], the major challenges are process integration related: improved strain engineering, achievement of lower device parasitics and adoption to the next generation substrates (450mm wafers).

As shown in Fig. 7 earlier, even much more radical novel device architectures are explored for future CMOS devices, including tunnel FETs (TFETs), carbon nanotube- (CNT) and graphene-based devices. However, most of these innovations are still in a state of preliminary exploration and are not ready yet for ultra-large-scale-integration (ULSI) processes.

Therefore in this Section, we will discuss the two major innovations that might be introduced in the nearby future and are only a few hurdles away for USLI integration: high-mobility FinFET and gate-all-around nanowire (GAA-NW) devices.

### *1.4.1 High-mobility FinFETs*

The use of high-mobility materials has been anticipated for a long time already. Strong attention has been paid to SiGe and Ge and III-V semiconductor materials as high mobility candidates. The major advantage is that the characteristic resistance of carriers flowing through the channel is lower, because of their lower effective mass. As such, a transistor can produce a larger drive current. This can be expressed as follows:

$$J = qE\mu n \tag{1.1}$$

where $E$ is the lateral electric field in the channel, $q$ is the elementary electron charge, $n$ the amount of carriers and $\mu$ the carrier mobility, valid for transistors in the drift regime. A similar expression can be obtained for transistors in the (quasi-)ballistic regime:

$$J = qnv_{inj} \tag{1.2}$$

where the drift current is determined by the carrier source injection velocity $v_{inj}$. This injection velocity is also higher for carriers with a reduced effective mass.

Because of the very low *electron* effective mass of III-V materials such as GaAs, InGaAs, InAs and InP, and the very low *hole* effective mass of Ge, these materials are suited for high performance n- and pMOSFET applications respectively. If they are co-integrable, CMOS applications are enabled too.

Apart from their enhanced mobility, the use of alternative channel materials also has other very interesting advantages: the off-current, as depicted earlier in Fig. 16, can now be suppressed by band engineering, i.e. by confining the high mobility channel with the underlying silicon substrate [Hellings10].

The ITRS 2010 was predicting the introduction of Ge and III-V (InGaAs) channels in 2018, which is in between the 10 and 7 nm node, the latest available roadmap, ITRS 2013, rather aims at the 7 nm node [ITRS].

Because the $SiO_2$ lost is privileged role as gate dielectric after the introduction of the high-k, the Si can now be replaced by high mobility materials. This also means that alternative gate dielectrics will need to be selected and developed. For the integration of high mobility channels, suitable high-κ metal gate dielectrics with low interface trap density ($D_{IT}$), low bulk traps and leakage, unpinned Fermi level and low ohmic contact resistances are major challenges [Fleetwood08].

The *opportunity window* for these high-mobility materials only opens when the above criteria are fulfilled. Interestingly, this window is also finite at the other end of the scale: both the subthreshold swing for devices with gate lengths below 10nm increases, as the intrinsic carrier confinement in the inversion layer is lower for devices with low effective masses. This is shown in Fig. 16 and is mostly true for III-V materials.

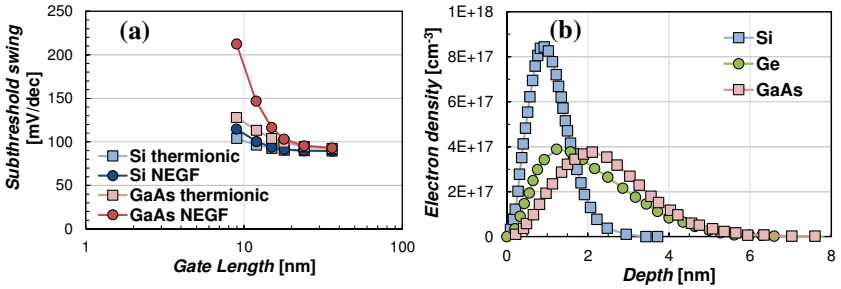



**Fig. 16: From (a) the characteristic subthreshold swing for various channel materials and (b) confinement of the inversion layer carriers as calculated with Schrödinger's equation it is clear that high-mobility materials are *less* scalable than silicon. Note that the simulations were performed on planar devices. [replotted from Smith07]**

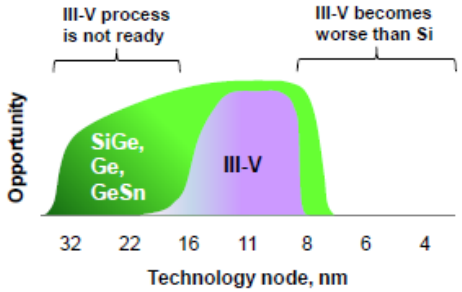



**Fig. 17: Schematic illustration of how** III-V **materials might arrive too late for use in ULSI applications. Note that the 14nm node is still using silicon channel material [replotted from Moroz11].**

Taking into account the reasoning that every new technology has to last for at least two nodes, it will depend on the speed of the developments in the upcoming years if these III-V materials will actually be introduced in ULSI applications, illustrated in Fig. 17. Moreover, device engineers have already experience with SiGe channel materials, which can be epitaxially grown on a Si lattice, this makes the SiGe (with high Ge content) or pure Ge as high mobility channels currently the most promising candidates for replacing Si pFETs in the 7 nm node.

### *1.4.2 Gate-all-around nanowire devices*

Nanowire based devices might be the next natural step in the scaling of micro-electronic circuits. The surrounding gate geometry exhibits similar properties and advantages as FinFET devices over planar devices. Even though it is not easy to realize such a device architecture with nanoscale features, they are compatible with both top-down as bottom-up fabrication approaches [Cui03, Goldman03]. In Fig. 18, a schematic image of a vertical and horizontal nanowire is given.

Typical *horizontal* silicon gate-all-around nanowire transistors (GAA NWFETs) have a layout similar to FinFET devices, but with a gate wrapped all around the confined channel. GAA-NWFETs have already been successfully characterized both theoretically and experimentally [Moroz14, Veloso15]. Due to their ultra-scaled dimensions with NW diameters below 10

nm, they offer superior electrostatics over FinFETs, reduced short channel effect, low leakage current and steep sub-threshold slope. Even though this makes them attractive for low-power applications, the drive current per NW is still in µA range. For that reason, multiple NW will have to be stacked to reach similar drive currents per footprint area as FinFET devices [Veloso15].

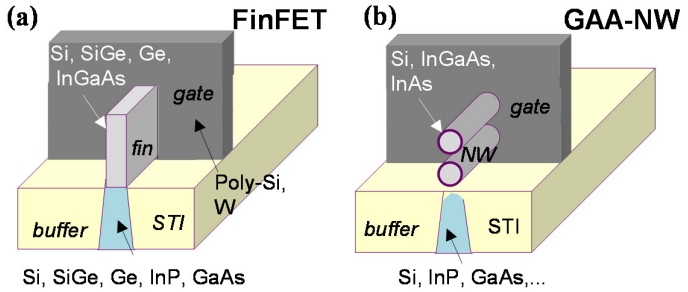



**Fig. 18: The geometric difference between (a) FinFET and (b) horizontal NWs is relatively small. The nanowires can be stacked to increase the drive current per footprint area.**

Interestingly, from this point of view, one might consider to increase the FinFET aspect ratio instead. However, not the fin *height*, but mostly the fin *width* has to shrink synchronously with the gate length to prevent excessive off-state leakage [Moroz14]. This means that the fin can potentially bend or collapse during manufacturing. The fin aspect ratio can thus not be ever-increased.

## 1.5 Reliability issues in scaling

Up to this point, we have paid little attention to r*eliability issues over several nodes*. Still, we can claim that these reliability issues *steered and influenced a lot of the design decisions in the past*. Already the very first transistor could only work after interface states were reduced with a surface cleaning treatment.

From that point on, silicon had a strong dominance in all the MOSFET products up to now, due to the existence of the excellent Si-$SiO_2$ interface, which plays a critical role in both the performance and reliability of the

device. For example, if the interface has many defects (the so called 'interface states'), or is rough, then the device's carrier mobility decreases, resulting in lower performance and a degraded long-term reliability [Sze06].

Also the quality of the gate oxide is critical. Its main goal is to prevent current from flowing between the gate and substrate electrodes. Both interface and oxide quality thus contribute to the performance of a MOSFETs to behave as an almost ideal switch, both in initial state (also called "time-zero") as throughout its lifetime.

Since the 70's, the electric fields, currents and power densities have only been increasing, except for the period of constant field scaling, as was shown in Fig. 4.The slackening of the EOT scaling recently, was a strong proof that device engineers came close to the limits of what is tolerable for reliable operation of the device. This was one of the reasons industry introduced the FinFET architecture: maintaining gate control without scaling the EOT. However, a reduced EOT—without a degraded oxide quality—would still be beneficial for the performance of FinFETs.

One of the typical degradation mechanisms that can degrade MOSFET performance is Bias Temperature Instability (BTI). BTI manifests itself as unwanted shifts of the MOSFET parameters over operation time [Nicollian71], i.e. as a form of time-dependent degradation [Fleetwood08]. It is typically accelerated over time by the gate oxide electric field and by temperature. Since the introduction of high-k gate dielectrics, this is the most critical problem to solve at this moment.

Time dependent dielectric breakdown (TDDB) and stress-induced-leakage-current (SILC) are other such typical degradation mechanisms, linked to defects in and nearby in the gate dielectric.

Finally, there is channel-hot-carrier (CHC) degradation. Hot carriers are particles that obtain high kinetic energy from being accelerated in the channel due to the high lateral electric field. These particles can be injected into the gate dielectric, where they can get trapped or cause interface states to be generated. In a way hot carrier degradation is thus similar as BTI. A major difference is that the former mechanism heat up the device, as these hot carriers also interact with lattice ions, thereby emitting phonons. This will

cause a localized heating in the device, so called device self-heating effect (SHE).

For a reliability engineer, the typical target is to safe-guard and guarantee a 10-year lifetime under nominal operating conditions of the transistors. As an example, the maximum safe gate voltage to achieve a 10 years lifetime for a certain failure mechanism—in this case TDDB—for stacks with different EOT's is shown in Fig. 19. Even though the TDDB margins are decreasing, this degradation mechanisms forms no acute problem in the recent nodes.

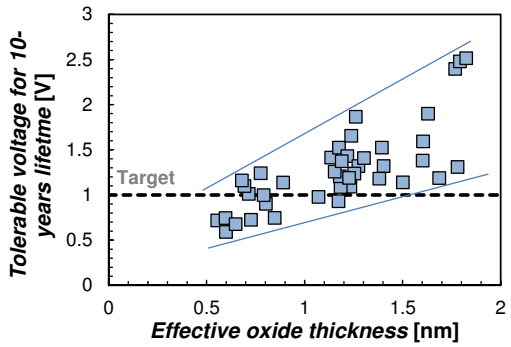



**Fig. 19: Ragnarsson et al. reported a TDDB lifetime of over 10 years with 0.5 nm EOT with a zero interface layer [Ragnar11].**

Concerning BTI degradation, the picture is vastly different. Since the introduction of SiON gate stacks, BTI has become a major concern for device engineers. In technologies using HK-MG, also strong positive BTI (PBTI) degradation has been reported for n-channel FETs. Especially for ultra-thin (UT) EOT devices an insufficient safe over-drive voltage is projected, as shown in Fig. 20. Even more worrying is that this trends appears to *accelerating* for EOTs below 1 nm. This degradation is even reported to increase exponentially with oxide thickness [Cartier11].

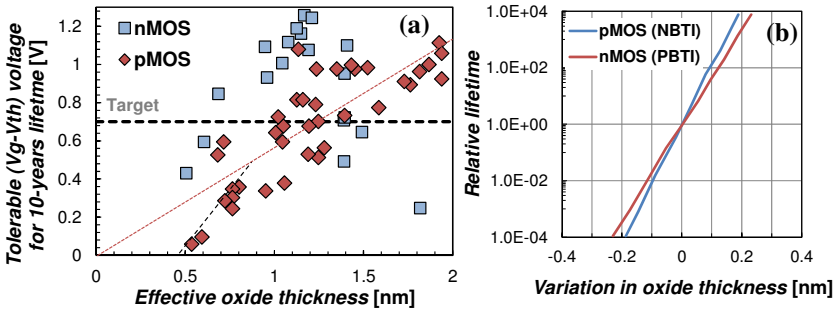



**Fig. 20: Trend of maximum tolerable overdrive ($V_G$-$V_{TH}$) gate voltage for a 10-years BTI lifetime showed a steep but linear decrease with EOT scaling. In the sub-1nm regime, this trend seems to be accelerated. (b) According to Cartier, this latter trend is even *exponential*. [replotted from Groes10 and Cartier11]**

## 1.6 Objectives of this thesis

As we study *major failure mechanisms for various devices types and architectures*, an overview of the known failure physics and models will be given in Chapter 2. The focus will be put on bias-temperature-instabilities (BTI), stress-induced-leakage currents (SILC), channel-hot-carrier degradation (CHC), time-dependent-dielectric-breakdown (TDDB), and device self-heating effects (SHE).

In Chapter 3, we discuss our developments to benchmark and to overcome the major device reliability hurdle for planar devices: BTI. We will provide *a physically-sound methodology* that allows *down-selecting gate stack materials and processes* suitable for ultra-thin EOT technologies *make refinements so that the best device performance* can be achieved.

In Chapter 4 we focus on the *understanding the nature of oxide charges and traps*, and how the different failure mechanisms can co-interact. We will observe how gate leakage currents and oxide charge trapping events are correlated. Newly observed properties regarding correlations in gate and drain current variations are discussed and a physical model is proposed that is capable of explaining the observed defect properties.

In Chapter 5, we elaborate on the device self-heating effects and their relation with the previously discussed degradation mechanisms. Various measurement methodologies will be discussed, and a new methodology based on nearby sensor-FET will be developed. Existing simulation methodologies will be applied and combined in multi-scale simulations and corroborated with dedicated measurements. This chapter will be accompanied with a case study of *SHE in planar, FinFET and gate-all-around nanowire (GAA-NW) devices*. We will discuss how the SHE can *impact device performance in circuits* and the *issues with reliability benchmarking*.

In Chapter 6, we elaborate on the self-heating effects in future device nodes that incorporate *novel channel materials*. We show that the presence of high-mobility materials and in particular alloys can bring along severe penalties on the thermal properties of the device.

## 1.7 References

[Arns98] Arns, R.G. "The other transistor: early history of the metal-oxide semiconductor field-effect transistor," IEEE Engineering Science and Education Journal. Vol. 7, Iss. 5, pp. 233-240, (1998).

[Auth12] Auth C. et al., "A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors 2012 Symposium on VLSI Technology", in Proc. VLSI Symposium on Technology, pp. 131-132, (2012)

[Bai04] Bai P., et al., "A 65nm Logic Technology Featuring 35nm Gate Lengths, Enhanced Channel Strain, 8 Cu Interconnect Layers, Low-κ ILD and 0.57μm2 SRAM Cell," in IEDM Tech. Dig., pp. 657-660, (2004).

[Baklan07] Baklanov M., Green M., and Maex K., "Dielectric Films for Advanced Microelectronics", Wiley, (2007).

[Bell48] Bell Telephone Laboratories — Technical Memorandum (May 28, 1948).

[Cartier11] Cartier, E. et al.,"Fundamental aspects of HfO2-based high-κ metal gate stack reliability and implications on tinv-scaling", in IEDM Tech. Dig., pp 18.4.1-18.4.4 , (2011).

[Chau04] Chau R. et al., "Advanced CMOS transistors in the nanotechnology era for high-performance, low-power logic applications", in IEEE Proc. ICSICT, pp. 26-30, (2004).

[Chidam04] Chidambaram P. R., et. al., "35% drive current improvement from recessed-SiGe drain extensions on 37 nm gate length PMOS", in Proc. VLSI Symposium on Technology, pp. 48-49, (2004).

[Choi12] Choi H-J., "Vapor–Liquid–Solid Growth of Semiconductor Nanowires", Chapter 1 in Springer Semiconductor Nanostructures for Optoelectronic Devices, Processing, Characterization and Applications, (2012).

[Colinge10] Colinge J.P. et al., "Nanowire transistors without junctions", Nature Nano. Vol. 5, No. 3, pp. 225-229, (2010).

[Colinge97] Colinge J.P., "Silicon-on-insulator technology: Materials to VLSI", 2nd Edition, Kluwer Academic Publishers, (1997).

[Compu15] Computer History Museum website, "1960 - Metal Oxide Semiconductor (MOS) Transistor Demonstrated", [accessed on 11/08/2015], online available: http://www.computerhistory.org/semiconductor/timeline/1960-MOS.html

[Cui03] Cui Y. et al., "High performance silicon nanowire field effect transistors", in Nano Lett. 3, 149, (2003).

[Dennard74] Dennard R., et al., "Design of ion-implanted MOSFETs with very small physical dimensions," IEEE Journal of Solid State Circuits, Vol. SC-9, No. 5, pp. 256-268, (1974).

[Doyle02] Doyle B. et al., "Transistor Elements for 30nm Physical Gate Length and Beyond", in Intel Technology Journ Vol 6, No. 2, pp. 42-54 (2002).

[Duan03] Duan X. et al., "High-performance thin-film transistors using semiconductor nanowires and nanoribbons", Nature, 425, 274, (2003).

[eetimes14] EE Times, 8/11/2014, "Intel Outlines 14nm, Broadwell", [accessed on 12/8/2015], URL:http://www.eetimes.com/document.asp?doc_id=1323476

[Fahad12] Fahad H.M and Hussain M. M., "Are Nanotube Architectures More Advantageous Than Nanowire Architectures For Field Effect Transistors?", Nature Scientific Report 475, Vol 2, (2012).

[Fleetwood08] Fleetwood D.M. , Defects in Microelectronic Materials and Devices.: CRC Press, (2008).

[Franco14] Franco J., Kaczer B. and Groeseneken G. "Reliability of High Mobility SiGe Channel MOSFETs for Future CMOS Applications", Springer, (2014)

[Ghani00] Ghani T. et al., "Scaling challenges and device design requirements for high performance sub-50 nm gate length planar CMOS transistors", in Proc. VLSI Symposium on Technology, pp. 174-175, (2000).

[Ghani03] Ghani T. et al., "A 90nm High Volume Manufacturing Logic Technology Featuring Novel 45nm Gate Length Strained Silicon CMOS Transistors," IEDM Tech. Dig. pp. 978-981, (2003).

[Goldberg06] Goldberger J. et al., "Silicon Vertically Integrated Nanowire Field Effect Transistors", in Nano Letters, Vol. 6, No. 5, pp. 973-977, (2006).

[Groes08] Groeseneken G. et al., "Reliability issues in MuGFET nanodevices", in Proc. IEEE International Reliability Physics Symposium, (IRPS), pp. 52-60, (2008).

[Groes10] Groeseneken G., Degraeve R., Kaczer B. and Martens K., "Trends and perspectives for electrical characterization and reliability assessment in advanced CMOS technologies", in Proc. ESSDERC, pp. 64-72, (2010).

[Hellings10] Hellings G. et al., "Implant-Free SiGe Quantum Well pFET: A novel, highly scalable and low thermal budget device, featuring raised source/drain and high-mobility channel", in IEDM Tech. Dig., pp. 241-244, (2010).

[Hou04] Hou T. et al., "Single Crystal Nanowire Vertical Surround-Gate Field-Effect Transistor", in Nano Letters 2004, Vol. 4, No. 7, pp. 1247-1252

[ITRS] International Technology Roadmap for Semiconductors (ITRS), online available at http://public.itrs.net

[Iwai09] Iwai H, "Roadmap for 22nm and beyond", Microelectronics Engineering, Vol. 86, Iss. 7–9, pp. 1520–1528, (2009).

[Jan09] Jan C. H. et al., "A 32nm SoC platform technology with 2nd generation high-k/metal gate transistors optimized for ultra low power, high performance, and high density product applications," in IEDM Tech. Dig., pp. 1-4, (2009).

[Kuhn09] Kuhn K. J., "Moore's Law Past 32nm: The Challenges in Physics and Technology Scaling", in 2009 International Conference on Solid State Devices and Materials (SSDM), Plenary t (2009).

[Lilienfeld28] Lilienfeld, J. E. "Device for controlling electric current," U. S. Patent No. 1,900,018, (1928).

[Mistry07] Mistry K., et al., "A 45nm Logic Technology with High-κ + Metal-Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging," in IEDM Tech. Dig., pp. 247-250, (2007).

[Moore65] Moore G. E., "Cramming more components onto integrated circuits", Electronics Magazine, vol. 38, pp. 4, (1965).

[Moore98] Moore, G. E. "The Role of Fairchild in Silicon Technology" Proceedings of the IEEE Vol. 86, Iss. 1 pp. 53-62, (1998).

[Moroz14] Moroz V. et al, "Modeling and optimization of group IV and III–V FinFETs and nano-wires", in IEDM Tech. Dig., pp 714-744 , (2014).

[Nara09] Nara Y, "Scaling Challenges of MOSFET for 32nm Node and Beyond", in Proc. VLSI Symposium on Technology, pp. 72-73, (2009).

[Nataraj14] Natarajan S. et al., "A 14nm Logic Technology Featuring 2$^{nd}$-Generation FinFET transistors, Air-Gapped Interconnects, Self-Aligned Double Patterning and a 0.0588μm² SRAM cell size", in IEDM Tech. Dig., pp. 71-73, (2004).

[Nicollian71] Nicollian E. et al., "Electrochemical charging of thermal SiO2 films injected electron currents", Journal of Applied Physics, Vol. 42, No. 13, pp. 5654-5664, (1971).

[Nowak02] Nowak E., "Maintaining the benefits of CMOS scaling when scaling bogs down", in IBM J. Res. & Dev., Vol. 46, No. 2-3, pp. 169-180, (2002).

[Park02] Park J.T. and Colinge J.P., "Muliple-gate SOI MOSFET's: device design guidelines", IEEE Trans. Electron Dev., vol. 49, pp. 2222-2229, (2002).

[Ragnar11] Ragnarsson, L-A et al., " Ultrathin EOT high-κ/metal gate devices for future technologies: Challenges, achievements and perspectives (invited)", in J. Microelectronic Engineering, Vol. 88, Iss. 7, pp. 1317-1322, (2011)

[Riordan99] Riordan M., Hoddeson L. and Herring C., "The invention of the transistor", Reviews of modern Physics, Vol. 71, No. 2, (1999).

[Schokley51] Shockley, W. "Circuit Element Utilizing Semiconductive Material," U. S. Patent 2,569,347, Filed June 26, 1948. Issued September 25, (1951).

[semimd12] Semiconductor Manufacturing & Design, 24/4/2012, "Intel's 22-nm Trigate Transistors Exposed", [accessed on 14/8/2015], available online: http://www.eetimes.com/document.asp?doc_id=1323476

[Smith07] Smith, L. et al., "Design Guidelines for High Mobility Channel Bulk n-MOSFETs", in Proc. Materials Research Society (MRS), Vol. 955, pp. 7, (2007).

[Sze06] Sze S.M. and Kwok K. Ng, "Physics of Semiconductor Devices", Wiley, 3rd Edition, 2006

[Thompson02] Thompson S. et al., "A 90 nm logic technology featuring 50 nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and 1µ/2µ/ SRAM cell", IEDM Tech Dig., pp. 61-64, (2002).

[Thompson04] Thompson S. et al., "A 90-nm Logic Technology Featuring Strained-Silicon", in Trans. Elec. Dev. Vol. 51, No. 11, pp. 1790-1797, (2004).

[Ti15] Texas Instruments website, "The chip that Jack built", [accessed on 14/8/2015], online available: http://www.ti.com/corp/docs/kilbyctr/jackbuilt.shtml

[Toleda11] Toledano-Luque M. et al., "From mean values to distributions of BTI lifetime of deeply scaled FETs through atomistic understanding of the degradation" in Proc. VLSI Symposium on Technology, pp. 152-153, (2011).

[Veloso11] Veloso A. et al., "Gate-last vs. gate-first technology for aggressively scaled EOT logic/RF CMOS", in Proc. VLSI Symposium on Technology, pp. 34-35, (2011).

[Veloso15] Veloso A. et al., "Gate-all-around NWFETs vs. triple-gate FinFETs: junctionless vx extensionless onventional junction devices with controlled EWF modulation for multi-VT CMOS", in Proc. VLSI Symposium on Technology, pp. 138-139, (2015).

# Chapter 2: Overview of failure mechanisms

This chapter will discuss the major failure mechanisms that are observed in MOSFET devices and that can cause reliability issues. Both their time-zero ("yield") as their time-dependent ("reliability") impact will be discussed.

## 2.1 Introduction

As discussed in Chapter 1, silicon dioxide ($SiO_2$) is the key material for the fabrication of stable and high performance MOS devices. The main motivation behind the use of silicon dioxide in MOS devices and integrated circuits are its unique properties: it is the only native oxide of a common semiconductor which is stable in water and at elevated temperatures, an excellent electrical insulator, a mask to common diffusing species, and capable of forming a nearly perfect electrical interface with its substrate [Dobkin03].

At present, $SiO_2$ layers can be produced with chargeable defect densities about $10^{10}$ $cm^{-2}$ and breakdown fields in excess of 10 MV/cm [Wallace2005]. These outstanding electrical properties clearly present a significant challenge for any alternative gate dielectric candidate.

Still, the study of optimizing $SiO_2$ has also been ongoing for a long time. The earliest attempts to fabricate MOSFET devices by Atalla and Kahng were unsuccessful because of the high defectivity of the oxide layers. In the early 60's, the various charges that are associated with the thermal oxidized silicon structure were described and linked with serious yield and reliability problems [Bentarzi11]. It is at that time that, as now generally established, four types of oxide charges were typically found in a MOS system: interface-trapped charge, fixed oxide charge, oxide-trapped charge, and mobile-ionic charge.

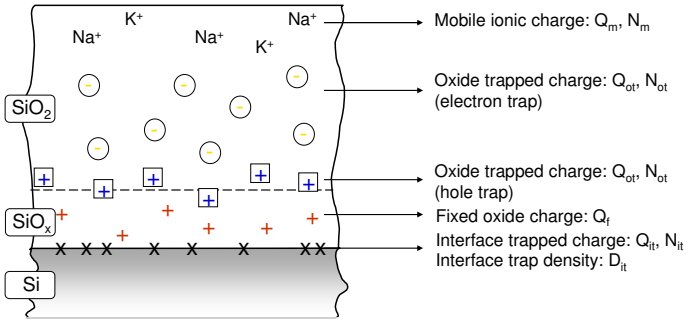



**Fig. 21: The various types of oxide charges that were identified in the '60s: mobile ionic charge, oxide trapped charge (electron and hole traps), fixed oxide charge, and interface trapped charges [replotted from Deal80].**

Except for Self-Heating Effects (SHE), which will be discussed at the end of this Chapter, other reliability issues that are discussed here, such as Bias Temperature Instabilities (BTI) and, related to that, Random-Telegraph Noise (RTN) are directly caused by charges or charge trapping in the dielectric interface or bulk oxide. Also Stress Induced Leakage Current (SILC) and Time-Dependent Dielectric Breakdown (TDDB) are linked to defect generation in the oxide, or at the interface, such as Channel Hot-Carrier degradation (CHC).

The overview that will be given here, is not aiming to completely cover the literature that is widely available around these topics, but is meant to give the reader a sufficient basis for the original work that is carried out by the author in the subsequent Chapters.

The sections below have similar structures: a phenomenological overview will be given, after which the basic interpretations and the physical background and current state-of-the-art models will be given. Finally, main observations and models are projected to nanoscale devices.

## 2.2 Bias Temperature Instabilities

Bias Temperature Instabilities are almost as old as metal-semiconductor-oxide (MOS) technology. One of the first sources referring to these bias temperature instabilities is found in '65 [Snow65], so only a few years after

Athalla and Kang's invention in '59 [Arns98]. Snow attributed these instabilities to the kinetics of alkali ions, migrating in thermally grown silicon dioxide films. He studied their transport through the oxide as a function of time, temperature, and applied voltage.

Deal reported an increase of surface state density during negative bias stress, and a recovery of these states under positive bias [Deal67]. It was Nicollian in 1971 who studied the impact of water on the $SiO_2$ formation who introduced the concept of an electrochemical reaction for the first time. In his description, water related defect centers could capture an electron and release a hydrogen atom. The latter would diffuse away and leave behind a fixed charge [Nicollian71]. This concept was concretized by Jeppson *et al*. who developed a physical model based on the breaking of Si-H bonds, forming dangling silicon bond defects and the diffusion of the charged H-related species into the oxide [Jeppson77].

However, Bias Temperature Instabilities remained a relatively unimportant phenomenon until the introduction of nitrogen in the $SiO_2$ in the early 2000's, as described in Chapter 1. Since then, the awareness for *negative* BTI in pMOSFETs grew. It was with the introduction of the high-k dielectrics in the 45nm node that also *positive* BTI gained similar amounts of interest.

As will become clear from the subsequent Sections, the interpretation of BTI has been a highly controversial subject since the 60's. Many observations, explanations and physical models are found in literature and are not always congruent.

### *2.2.1 Phenomenological overview*

In a typical CMOS logical circuit, nMOS and pMOS devices each have their proper function; the pMOS works as a pull-up, whereas the nMOS works as a pull down. In the passive state of the circuit, i.e. when it is not switching, a constant bias is applied to the gates, as depicted in Fig. 22. Depending on the input of the circuit either the nMOS or the pMOS are biased into the inversion mode.

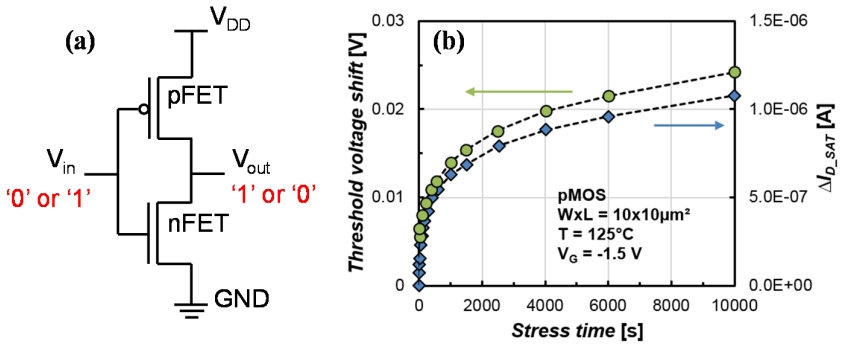



**Fig. 22: (a) Typical bias conditions of a p- and nMOS devices in an inverter circuit and (b) change in threshold voltage and saturation drive current characteristics of a pMOS device in stress conditions.**

As the name suggests, BTI refers to a time-dependent instability in transistors, accelerated with gate voltage (or '*bias*') and *temperature*. The instability is best represented as a shift of the threshold voltage ($V_{TH}$) or as a reduction of the saturation drive current ($I_{D\_SAT}$). This reduction in drive current will ultimately result in a speed reduction of the circuit (or indirectly by reduced margins, such as static noise margin for SRAM cells). For p-channel MOSFETs the term negative bias temperature instability (NBTI) is used, whereas for n-channel MOSFETs the degradation is called positive bias temperature instability (PBTI) since the corresponding gate bias conditions are negative and positive, respectively.

Even though there are distinct differences in the physical degradation mechanisms—mainly due to the asymmetry in the band and the defect structure for high-k MOSFETs, characterization methods to study NBTI and can also be used to study PBTI. In Fig. 23, the schematic band structures of the p-channel and n-channel MOSFETs with MG/HK stacks during operation (i.e. in inversion) are compared, and the dominant charge trapping effects are indicated.

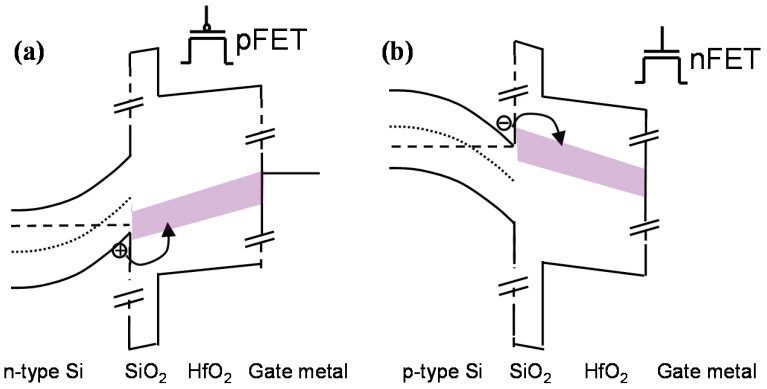



**Fig. 23: Band diagrams of high-k/metal gate (a) nMOSFET and (b) pMOSFET biased in inversion. Note the distinct difference in the physical and energetic locations of the charge traps for both cases.**

The schematic in Fig. 23 also shows the distinct difference between the charge trap locations for N and PBTI. For PBTI, the charges are mostly trapped in the high-k bulk and the $SiO_2$/high-k interface, i.e. away from the channel. Therefore, the Coulombic scattering interaction between the trapped charges and the carriers in the channel inversion layer carriers is low. As a result, the carrier mobility is less impacted by this trap mechanism and the result is only a $V_{TH}$ shift. For NBTI, charges can also be trapped at or near the Si-$SiO_2$ interface, thus influencing the mobility. The result is visible a reduction in the transconductance ($g_m$) peak and a degradation of the subthreshold swing (SS) [Mitani05], [Garros10], as shown in Fig. 24.

Another typical feature that is observed in BTI is the partial recovery or 'relaxation' of the degradation, once the stress voltage is removed. This means that charges trapped in the oxide are released again in the channel. Even though this recovery might seem to be a virtue, it actually impedes the correct quantification of the '*real' $V_{TH}$* shift. To understand these charge de-trapping kinetics, numerous efforts have been done to study this BTI relaxation, as will be discussed later in this Section.

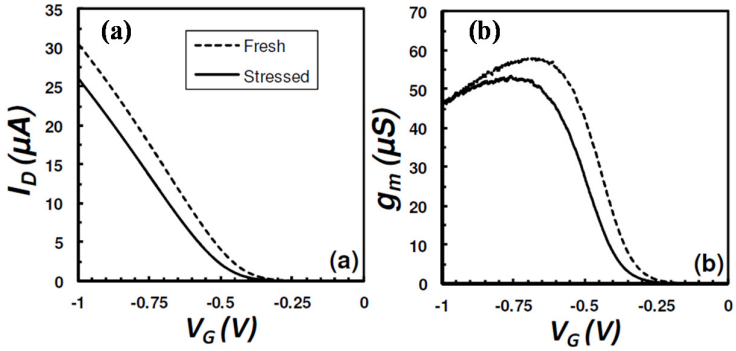



**Fig. 24: Change in current characteristics of a pMOS after NBTI stress. Apart from the horizontal translation ($V_{TH}$ shift) of the $I_D$-$V_G$ characteristic in (a), a drop in transconductance is also observed in (b). [replotted from Franco12]**

## 2.2.2 Empirical modeling of BTI

A typical example of evaluating the BTI is the measure-stress-measure (MSM) procedure as depicted in Fig. 25 below. In this case, the $\Delta V_{TH}$ is evaluated periodically, after every stress cycle, by evaluation of the transistor $I_D$-$V_G$. Typically, the accumulated stress time is exponentially increased, as it is known that the $\Delta V_{TH}$ follows a power-law as a function of accumulated stress time ($t_s$) within the typical measurement windows:

$$\Delta V_{TH} = A t_s^{\,n} \tag{2.1}$$

with $A$ being a stress voltage dependent pre-factor and $n$ the power-law time exponent—typically between 0.1-0.25—subject for controversial interpretations as will be discussed in Section 2.2.10.

Also, it should be noted that recently it was shown by Franco *et al.* that this power law is a simplification and by rescaling the $\Delta V_{TH}$ traces recorded at different overdrive voltages $V_{ov}$ along the time axis, a universal curve could be obtained, which clearly showed a *saturating behavior* (which the power law curve will not show). This experimental curve was thereafter shown to be

perfectly described by a Cumulative Lognormal distribution, rather than a power law [Franco14].

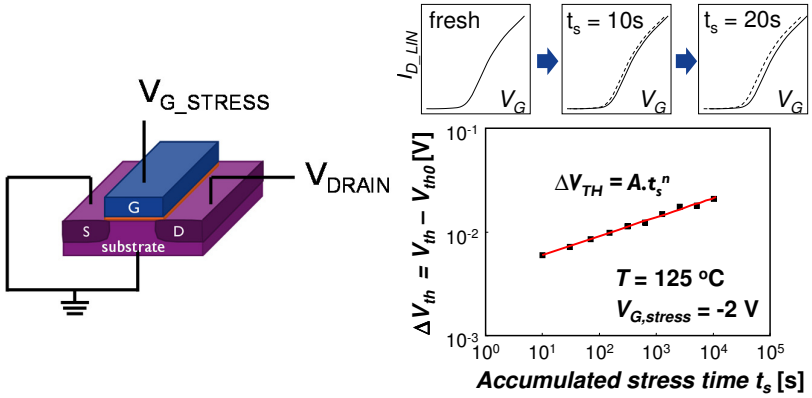



**Fig. 25: The shift of the threshold voltage ($\Delta V_{TH}$) can be evaluated after every stress phase. The $\Delta V_{TH}$ appears to increase according to a power-law [replotted from Kaczer08].**

By expanding the measurement set for various stress bias conditions, the voltage dependence of the stress-dependent pre-factor $A$ can also be derived. The result typically looks as illustrated in Fig. 26, and the pre-factor is shown to follow a power-law dependence *itself* with respect to the effective electric field $E_{OX}$ applied on the gate stack:

$$A \propto E_{ox}^{\gamma} \approx \left( \frac{V_{ov}}{t_{ox}} \right)^{\gamma} \tag{2.2}$$

with $V_{OV}$ being the overdrive voltage, $t_{ox}$ the oxide equivalent thickness and $\gamma$ being the voltage-exponent. The latter is typically around 3 for devices with a silicon substrate and a $SiO_x$-based interfacial layer (IL) for the gate dielectric during negative bias stress [Franco13]. The time-dependence of the power law, $n$, remains rather constant for these test conditions (see Fig. 26(a)).

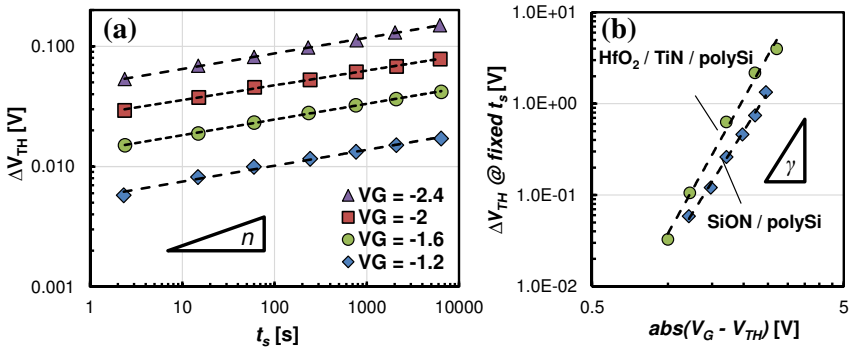



**Fig. 26: (a) The $\Delta V_{TH}$ increases with stress voltage but the power law time constant (i.e. the slope) remains mostly unchanged. (b) The voltage acceleration can be depicted at a fixed stress condition. The latter is shown to be dependent on the overdrive voltage and follows a power law dependence *itself* with exponents γ around 3 for Si/SiO₂ IL-based systems [replotted from Franco and Kerber08].**

Finally, the temperature dependence of the BTI-induced $V_{TH}$ degradation is commonly accepted to follow the Arrhenius law:

$$\Delta V_{TH} \propto e^{\left(\frac{-E_A}{k_b T}\right)} \tag{2.3}$$

with $E_a$ the *apparent* activation energy, $k_b$ the Boltzmann constant and $T$ the temperature during the measurements. The activation energy *for the* $\Delta V_{TH}$ is typically 60~80meV [Huard06], shown in Fig. 27. The 'real' activation energy is defined as *Ea/n*. Note that the activation energy *for the extracted device lifetimes* based upon a certain failure criterion differs from this one.

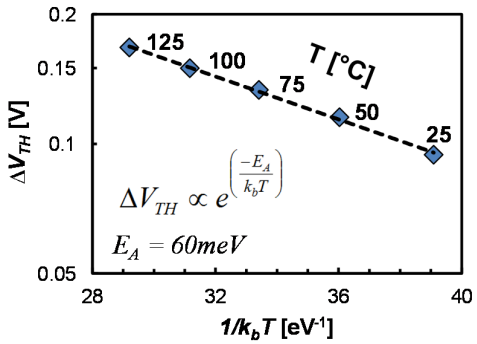



**Fig. 27: The impact of temperature on the $\Delta V_{TH}$ is shown to follow the Arrhenius law. [replotted from Franco12]**

As all the known dependencies—stress *field*, *time* and *temperature*—for BTI degradation have been characterized, the above dependencies can be summarized in the following equation:

$$\Delta V_{TH} \approx C.t_s^n.e^{\left(\frac{|V_G - V_{TH0}|}{t_{ox}}\right)^{\gamma}\left(\frac{-E_A}{k_bT}\right)} \quad . \qquad (2.4)$$

By itself, the above *empirical* formula would be sufficient to *back*-extrapolate the device threshold voltage shift *towards realistic operating conditions*, in the assumption that the proposed relations stay valid at lower fields, temperatures and extended time scales. Typically, a failure criterion such as $\Delta V_{TH}$ = 30~50mV is used. Such extrapolations with power laws always exhibit a significant risk as a small deviation in one of the exponents, such as the time exponent *n*, can easily cause the outcome of the extrapolation to change dramatically.

### 2.2.3 Recovery of $V_{TH}$

We did not yet take into account the *temporal evolution* of the $\Delta V_{TH}$ shift *after stress*. With the introduction of SiON and high-k dielectrics, it was shown that the $V_{TH}$ degradation is partially recoverable [Ershov03]. Immediately after releasing the stress voltage, the $\Delta V_{TH}$ is reduced. The typical

time range of this relaxation spans over *all decades observable in laboratory environment*, i.e. from below micro-seconds up to hours and days [Kerber08]. This relaxation phenomenon has profound implications on understanding the dynamics of BTI and how these measured effects should be propagated to real circuit level. For example, as a result of this relaxation effect, a transistor under AC-stress will experience a lower degradation than its counterpart under equivalent DC stress. Detailed study of this relaxation mechanisms can help to understand the underlying mechanism that governs the BTI kinetics. The time dependence of the $V_{TH}$ during stress and subsequent relaxation is depicted in Fig. 28 on a linear and semi-log scale.

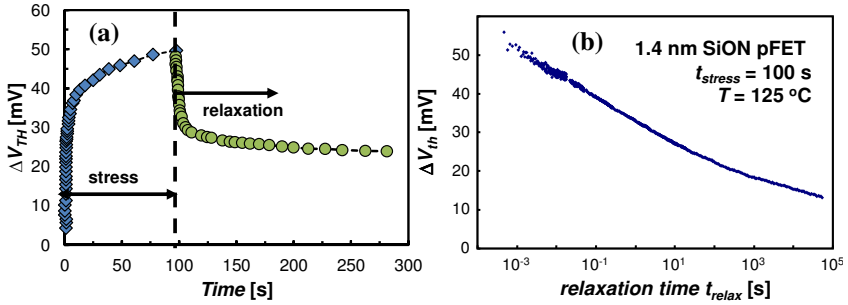



**Fig. 28: (a) Time trace of $V_{TH}$ shift during BTI stress and subsequent relaxation. (b) Relaxation of the $\Delta V_{TH}$ spans over multiple decades of time. The relaxation starts well *below* 1ms. The uncertainty of the actual $\Delta V_{TH}$ propogates into questionable extrapolations. [replotted from Kaczer08]**

As seen in Fig. 28, the start of the $\Delta V_{TH}$ relaxation cannot be easily captured. The actual $V_{TH}$ shift that is measured will thus depend on the measurement equipment and configuration that is used. It is clear that such an arbitrary choice for the $V_{TH}$ relaxation time cannot guarantee correct device lifetime extrapolations.

### *2.2.4 Universal recovery model for $V_{TH}$*

Primarily, [Grasser07] and [Kaczer08] tried to capture the recovery kinetics using the universal recovery behavior. A measurement procedure was

proposed by Kaczer *et al*. to study the $\Delta V_{TH}$ relaxation: the extended measure-stress-measure (eMSM) technique [Kaczer08]. In this technique, short recordings of the relaxation during each measurement phase allow monitoring and correcting for the otherwise-unknown relaxation component.

The major advantages of eMSM include discerning artefacts and anomalous BTI due to bipolar trapping, otherwise concealed to other techniques. In the *I-V*-MSM sweep, the four terminals of the device-under-test (DUT) are connected to Keithley 2602 or 2636 source measurement units (SMUs). The DUT is biased at the drain in the linear regime.

Fig. 29 schematically depicts the proposed scheme, achievable with conventional equipment. The extracted $V_{TH}$ transients are depicted in Fig. 30.

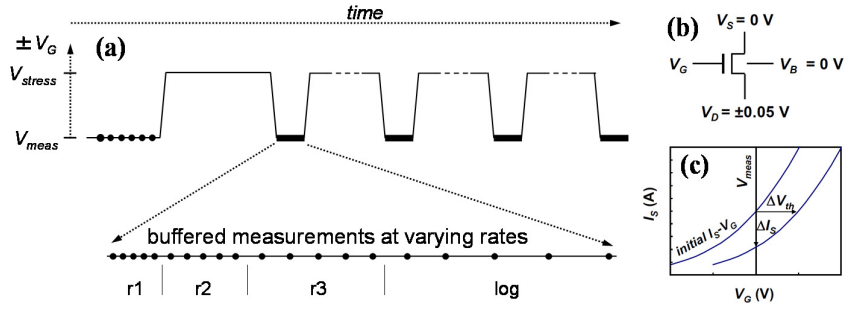



**Fig. 29: (a) Principle of the extended Measure-Stress-Measure (eMSM) technique. $\Delta V_{TH}$ transients are measured for multiple decades after every stress cycle. "r1", "r2", "r3' and "log" represent different sampling rates to efficiently cover the logarithmic time scale. (b) Bias applied on the DUT during the measurement and (c) illustration of the FET current-to-$\Delta V_{TH}$ conversion [replotted from Kaczer08].**

The threshold voltage shift is attributed to two components, a permanent component '*P*' and a recoverable component '*R*' [Kaczer08]:

$$\Delta V_{TH} = R(t_{recovery}) + P(t_{stress}) \quad . \tag{2.5}$$

The permanent component is the component that is not prone to any relaxation, whereas the recoverable component can be extrapolated to infinitely short relaxation times (i.e. $t_{RELAX} = 0$). These extrapolations can be made because of the existence of a *universal recovery curve*, on which the relaxation data at the various stress times coincide, and is described as follows:

$$r(\xi) = \frac{\Delta V_{TH}(t_s)}{\Delta V_{TH}(t_r)} = \frac{1}{1 + B.\xi^{\beta}} \quad . \tag{2.6}$$

with $\Delta V_{TH}(t_r)$ and $\Delta V_{TH}(t_s)$ being the shifts at the beginning and during the recovery, B a fitting parameter and β approximately the inverse of the power law time exponent attributed to the dispersive transport properties, and ξ the universal relaxation time, defined as

$$\xi = \frac{t_s}{t_r} \quad . \tag{2.7}$$

The total $V_{TH}$ shift at any moment of the relaxation can then be defined as:

$$\Delta V_{TH} = R(t_{stress}, t_{relax}) + P(t_{stress}) = \frac{R(t_{relax})}{1 + B.\xi^{\beta}} + P(t_{stress}) \quad . \tag{2.8}$$

An example of the fitted recovery data using one ξ is shown in Fig. 31.

Concluding, the measured relaxation traces can be fitted with the above described empirical model in order to estimate the full NBTI degradation as if it was measured directly after stress removal (i.e., with zero delay). Moreover, this formula allows to reconstruct the degradation that would be measured if the degradation would be measured with any other technique. This is illustrated in Fig. 32.

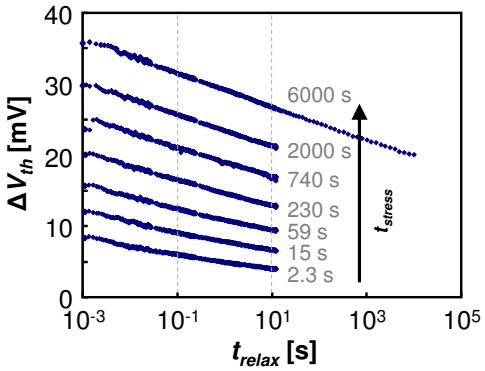



**Fig. 30: Example of $\Delta V_{TH}$ transients measured after every stress cycle. The cumulative stress times are exponentially increasing, yielding approximately linearly increasing $\Delta V_{TH}$ transients.**

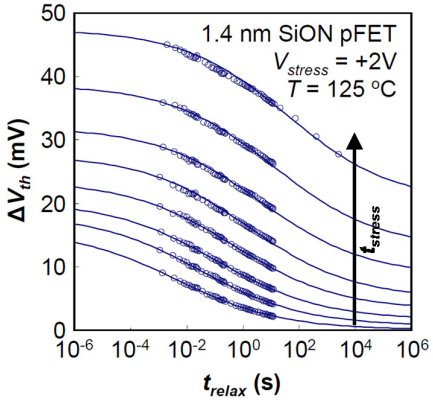



**Fig. 31: $\Delta V_{TH}$ recovery as a function of the relaxation time for increasing stress times, now fitted using the universal recovery behavior.**

The above described technique, i.e. fitting the relaxation data with a universal relaxation curve, only works for gate stacks that have a single defect band. It was shown that La- or Y-doped gate stacks tend to show defect bands that charge with different kinetics and voltage dependencies [Kaczer08]. In those cases, the $\Delta V_{TH}$ does not tend to relax monotonically. Finally, it has to be stated that other degradation parameters of the device, such as subthreshold

swing degradation or mobility reduction remain concealed with this technique.

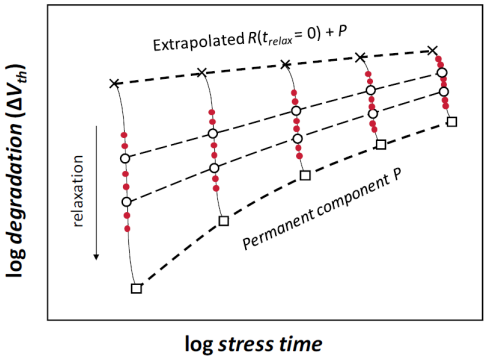



**Fig. 32: Illustration of how data points can be constructed as if the degradation would be measured on-the-fly or at any random given relaxation time utilizing the universal relaxation model.**

## 2.2.5 Lifetime extrapolation and benchmarking

In semiconductor industry, one is interested to compare and benchmark the BTI reliability of various gate stacks. To perform these gate stack quality benchmarks, a criterion such as 50mV of threshold voltage shift typically used to represent failure of the device. The time before the device exceeds this threshold voltage criterion is called the device lifetime.

The procedure to extract the lifetime is as follows. The device-under-test or 'DUT' is stressed at a certain 'overdrive' voltage ($V_{OV}$) with an eMSM scheme. The overdrive voltage corresponds to the gate bias exceeding the $V_{TH}$ ($V_G$-$V_{TH}$). It should be noted that this $V_{OV}$ is an indirect approximation for the electric field, as will be discussed in more detail in Section 3.5 A pristine device is selected for every $V_{OV}$. The temperature during the test is 125°C. After every stress phase of the eMSM, the $\Delta V_{TH}$ is systematically evaluated at 1ms of relaxation, which corresponds to the minimum delay time needed with a typical SMU. The time-to-failure (i.e. when the device reaches the critical $V_{TH}$ shift) can then be intra- or extrapolated with the empirical formula given in Eq. 2.1. Preferentially, the $V_{OV}$ is chosen such that the degradation criterion

is met during the time-window of the test. In this way, we avoid large extrapolations, sensitive to the time-exponent *n* (Fig. 33).

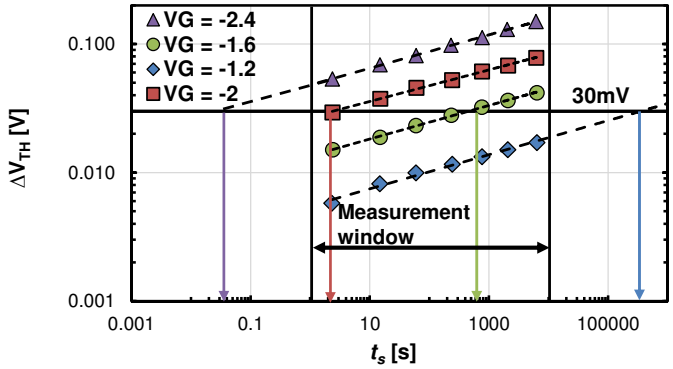



**Fig. 33: Principle of lifetime extrapolations of identical DUTs (device-under-test) at various stress conditions. Ideally, the device reaches the failure criterion within the measurement window, to avoid extrapolations sensitive to the time-exponent *n*, as described in Eq. 2.1.**

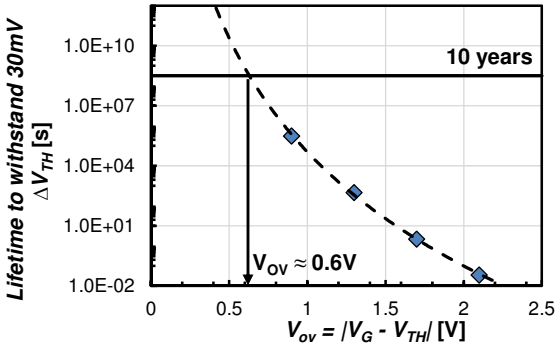



**Fig. 34: The maximum operating overdrive voltage for continuous 10 years operation can be estimated by plotting the extracted time-to-failure for various $V_{OV}$ and fitting the data with a power law for devices with EOT > 1nm. For devices with < 1nm EOT, an exponential function typically shows a better fit.**

The 10-year lifetime operating $V_{OV}$ can then be extrapolated given the dataset acquired above, now depicted in Fig. 34. Typically, the extrapolation

is performed using a power-law, which is found to fit the data with EOT > 1nm best. For UT-EOT devices, exponential dependences are often better suited.

### 2.2.6 Discussing the failure criterion

A few important side nodes have to be made when considering these lifetime extrapolations. A criterion such as 30 or 50mV of threshold voltage shift is typically used to represent failure of the device. Even though the actual origin of this criterion is unknown, this amount of threshold voltage shift tended to agree with a drive current degradation of roughly 5-10% in sub-100nm technology nodes.

The *arbitrariness* of this criterion however can easily be understood given the fact that the $V_{DD}$ and $V_{TH}$ are scaling at a difference pace from node to node. Novel devices operate closer to the $V_{TH}$, which increases the relative impact of a fixed $\Delta V_{TH}$ on the drive current. The increasing *relative* impacts of a 50mV $\Delta V_{TH}$ extracted for a few recent technology nodes is depicted in Fig. 35.

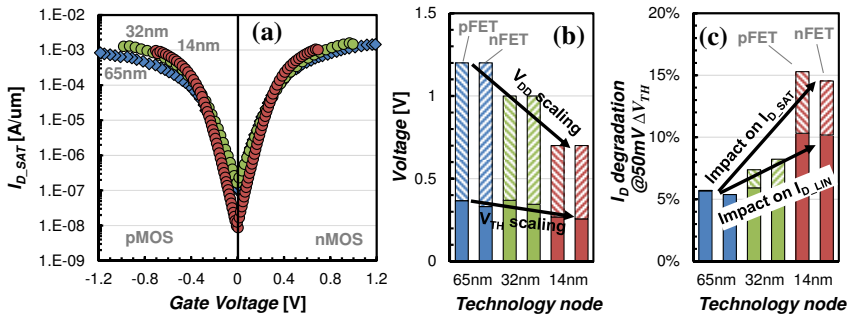



**Fig. 35: (a) Transistor saturation drive currents at their respective $V_{DD}$'s for various technology nodes. From (b) it is clear that device $V_{TH}$ does not scale according to the $V_{DD}$. (c) The impact of 50 mV $\Delta V_{TH}$ on the drive current is increasing from node to node, especially the saturation drain current, indicating the irrelevance of a fixed $V_{TH}$ shift criterion on a technology level. [$I_D$-$V_G$'s replotted from Auth09, Auth12 and Natarajan14]**

It can be argued that the failure criterion should in fact be adapted towards the technology for which the gate dielectric stack is foreseen to be used. This is important because any failure criterion itself incorporates an "arbitrary" mix of the BTI *kinetics* and *bias dependence* in the respective extrapolations that are performed to obtain the lifetime, as explained above.

Still, using a fixed $\Delta V_{TH}$ failure criterion is thought to be the most convenient tool to assess the gate stack quality and for benchmarking—albeit if the above considerations are taken into account—even though it is not always directly relevant on an application level. The latter fact will even become more apparent when we look into nanoscale devices, in the following Section.

### 2.2.7 Observations in nanoscale devices - variability

BTI in nanoscale devices is often referred to as 'time-dependent *variability*'. Indeed, one of the most important consequences of scaling devices into nanometer dimensions, is that nominally identically devices will no longer exhibit identical electrical characteristics. To understand the concept of 'time-dependent variability' in nanoscale devices, we briefly introduce the concept of 'variability' and the origin of the phenomenon.

Device-to-device variability can be categorized into extrinsic and intrinsic sources. We suggest the following categorization: *extrinsic* variability is related to the processing of the device, for example the gate dielectric thickness can vary. This will have an impact on the channel control and thus impact the electrical properties of the device. Devices that exhibit extrinsic variability should not be categorized as nominally identical, as these variability sources can be reduced by process optimizations.

*Intrinsic* variability has also an origin in the processing of the device, but is assumed to be non-controllable, i.e., process optimizations will not reduce these variability sources. Typically, these variability sources are related to the near-atomic dimensions of nanoscale devices, illustrated in Fig. 36. Such intrinsic sources can be line-edge-roughness (LER) or metal gate granularity (MGG) [Asenov08]. Also defects in the gate oxide nearby critical locations in the channel can significantly impact the device current. The nature of these intrinsic variability sources implicates that *these* device-to-device variations

will only increase as the devices shrink. In practice however, it is not always straightforward to distinguish if the observed variability is due to an extrinsic or intrinsic source. Using the matched pairs principle, the *extrinsic* or *systematic* variability can be separated from the *intrinsic or random* variability [Kaczer15, Giles15].

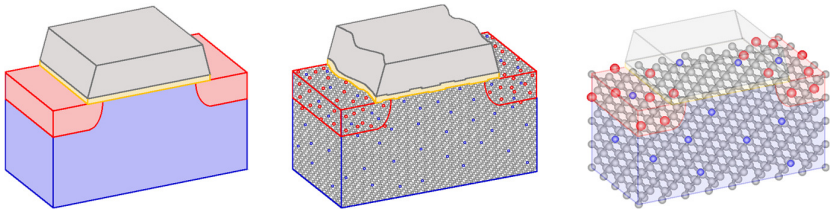


**Fig. 36: Whereas devices could be treated as consisting of continious bulk material in earlier technology nodes, their atomistic nature should now be taken into account as devices consist of a numerable amount of atoms. [replotted from Asenov08]**

### 2.2.8 Time-0 variability

The variability of the devices is typically expressed by means of the $V_{TH}$ distribution (Fig. 37(a)). The total time-zero variability can typically be well described with a Gaussian distribution [Kerber13, Giles15], as the major sources of variability typically exhibit normal distribution themselves. As can be seen in Fig. 37(b), also the time-zero *random* variability obtained by measuring tens of millions of devices, behaves excellently Gaussian (i.e. linear on a probit plot), but is *reduced* when moving from the 22nm towards the 14nm node in Intel's technology. This seems contradictory given the above explanation, but can be explained with Pelgrom's law.

In 1989, Pelgrom proposed and demonstrated the simple evaluation method of this random variation [Pelgrom89]. This is based on simple statistics and very useful: the so-called Pelgrom plot [Fig. 38(a)] can predict the $V_{TH}$ variation for scaled devices. The time-zero $V_{TH}$ variability is expected to scale inversely with the device channel area, and proportionally with the inversion layer thickness $T_{inv}$ and with the square root of the dopant concentration $N_{SUB}$ and the depletion depth $W_{dep}$ [Takeuchi07].

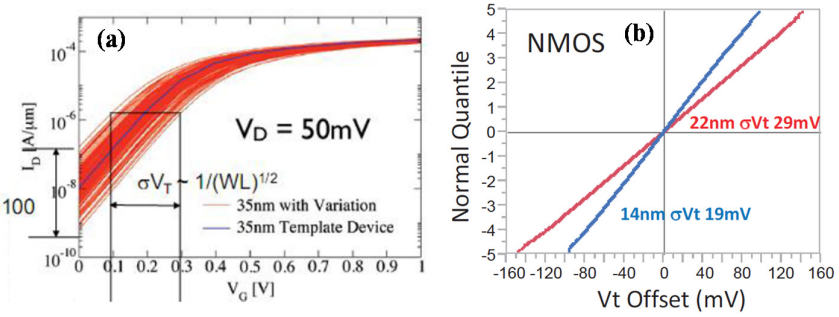



**Fig. 37: (a) Device-to-device variability is represented as a distribution of the $V_{TH}$. (b) A *reduction* of the $V_{TH}$ variability, represented by the slope of the distribution, is seen moving from the 22nm towards the 14nm node, indicating that the variability sources for *random* variability were reduced [replotted from Asenov08 and Giles15].**

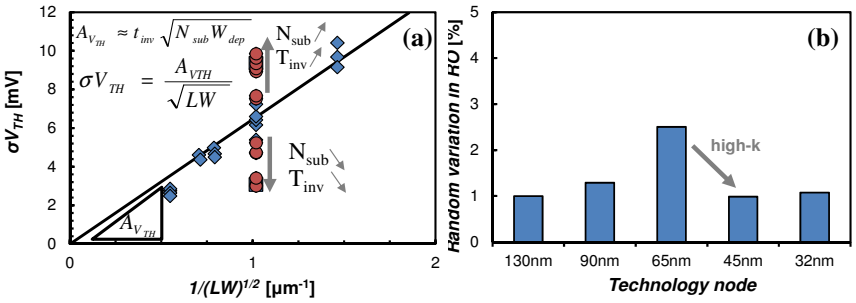



**Fig. 38: (a) Pelgrom plot representing the increasing random device-to-device variability with decreasing device area, for devices with *identical* intrinsic properties. The variability can however be increased or decreased depending on the doping level and inversion layer thickness. (b) This principle was brought in practice as Intel strongly reduced the EOT with the introduction of high-k dielectrics in the 45nm node [replotted from Takeuchi07 and Kuhn09].**

Both technology nodes in Fig. 37(b) utilize FinFET devices, the remarkable reduction in this $V_{TH}$-spread thus cannot be reduced by lowering doping concentrations in this case as the FinFETs are already un-doped and fully depleted in the 22nm node. This suggests that either the *intrinsic* variability sources for variability were reduced by reducing the gate oxide. This effect

was also seen before when industry moved from SiON to a high-k technology nodes, implicating a strong reduction in the EOT, as shown in Fig. 38(b). However, as the EOT reduction from the 22nm up to the 14nm node in Fig. 37 is most likely rather small, a more probable option is that the variability was in this case reduced by *extrinsic* factors, for example by changing the amorphisation degree of the metal gate (i.e. modifying its granularity).

### 2.2.9 Time-dependent variability

After we have described time-zero variability, we can proceed to the concept of *time-dependent*—in this case BTI-induced—variability. To understand the time-dependent scaling trends, we have to understand the source of these variations. When studying BTI in deeply-scaled devices instead of their large-area counterparts (i.e. those that are typically used for NBTI testing), one crucial observation was made that attracted a lot of attention: the recovery in small-area devices proceeds in *discrete* steps. Each individual step is believed to be caused by the discharging of an individual defect in the oxide [Grasser10a, Kaczer10]. This is depicted in Fig. 39.

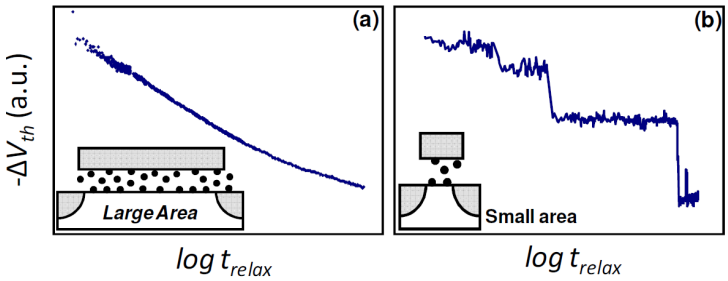



**Fig. 39: (a) Relaxation traces following a smooth relaxation over logarithmically increasing relaxation time in large area devices (b) are found to behave step-like in small-area devices. The large number of discharging defects in large-area devices is causing the smooth transient, whereas only a limited number of defects are visible in nanoscale devices.**

It was also shown by Grasser, overlaying the relaxation properties of multiple small devices, that each component exists of a particular sub-set of a few oxide defects, but originating from the same defect distribution as that

found in large-area devices. Based on this interpretation, it can be seen that the charging of individual defects is averaged out in a large device, whereas it will yield a strong statistical variation in small devices with identical defect density, as illustrated in the simulations in Fig. 40.

This has a major impact on how device lifetime extrapolations should be interpreted for nanoscale devices. In order to quantify this effect, one has to understand the underlying defect properties.

It was found that $V_{TH}$ shifts induced by defects follow an exponential distribution in terms of their impact [Kaczer10]. The probability-density-function (PDF) of a defect exhibiting an impact of $\Delta V_{TH}$ is described as:

$$f_1(\Delta V_{TH}) = \frac{1}{\eta} e^{\frac{\Delta V_{TH}}{\eta}} \tag{2.9}$$

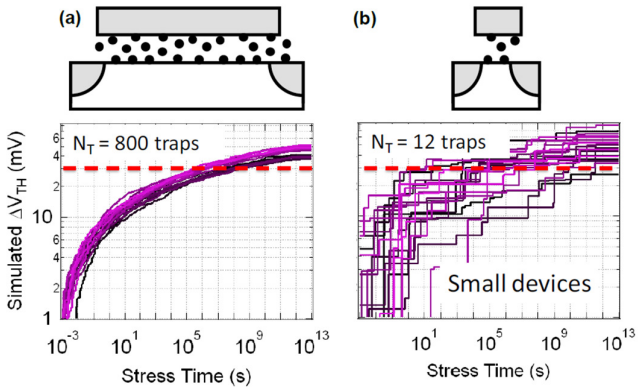



**Fig. 40: (a) In a large device the individual defect properties are averaged out (b) whereas in a small device a few defects with stochastically distributed properties control the degradation. The variation of the overall degradation for identical workloads will thus be larger in the latter [replotted from Grasser10a].**

with $\eta$ the mean impact per defect on the $V_{TH}$. As such, defects can have a $\Delta V_{TH}$ much larger than the value one would expect based on the simple charge sheet approximation in Eq. 2.8. Even a single charged defect can cause up to

tens of mV of $\Delta V_{TH}$ [Toledano11]. This means that trapping of one single "unlucky" charge can degrade the device up to the failure criterion.

The exponential distribution of single-charge $\Delta V_{TH}$ can be understood if non-uniformities in the channel of the FET due to dopant fluctuations are considered. The $V_{TH}$ of a device corresponds to the formation of a percolation path in the potential surface (mainly formed due to the random distribution of the dopants) between the device's source and drain. Depending on the position of the oxide charge, the percolation path can be affected by the defect when charged. The drop in the current then has to be compensated by an increase of the gate voltage, resulting in the observed $\Delta V_{TH}$. This principle is illustrated in Fig. 41.

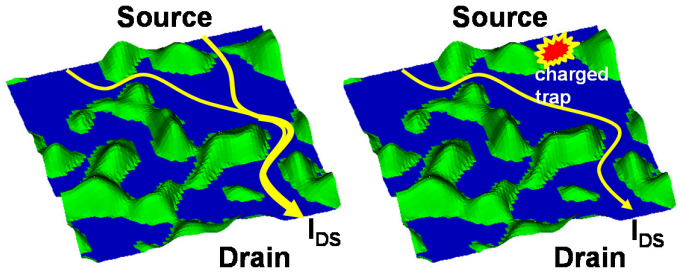



**Fig. 41: The random distribution of dopants gives rise to a potential profile in the channel of nanoscale devices. As a result, the current flows through percolation paths. If a trap in the gate oxide falls right on top of such a percolation path, it will have a dramatic impact on the total drive current of the transistor, thus exhibit a large $\Delta V_{TH}$.**

For multiple number of traps $N_T$, the total PDF can be written as a convolution of these exponential distributions as each of them is uncorrelated:

$$f_{N_T}(\Delta V_{TH}) = \frac{e^{-\frac{\Delta V_{TH}}{\eta}}}{(N_T - 1)!} \frac{\Delta V_{TH}^{N_T - 1}}{\eta^{N_T}} \quad , \tag{2.10}$$

and the corresponding cumulative density function is then:

$$F_N(\Delta V_{TH}) = 1 - \frac{\Gamma(N_T, \Delta V_{TH}/\eta)}{(N_T - 1)!} \tag{2.11}$$

with $\Gamma(N_T, \Delta V_{TH}/\eta)$ the mathematical gamma function.

The other major parameter describing the total $V_{TH}$ shift, apart from the impact per trap $\eta$, is the number of traps $N_T$. There are several possibilities to extract this number, either on a nanoscale device or by extrapolating it from the properties of a large-area device. However, the observations in nanoscale devices could be explained by [Kaczer11] if the number of defects per device $N$ was assumed to follow a Poisson distribution and was also already suggested by [Rauch07, Huard08]:

$$P_N(N) = \frac{e^{-N} N^{N_T}}{N_T!} \tag{2.12}$$

with $N_T$ is the mean number of defects in the gate oxide, directly related to the trap *density* $N_{OT}$.

Combining both distributions, i.e. summating the number of traps with the impact of each specific trap, yields the distribution of the *total* $V_{TH}$:

$$F_{N,\eta}(\Delta V_{TH}) = \sum_{N=1}^{\infty} \frac{e^{-N} N^{N_T}}{N_T!} F_{N_T}(\Delta V_{TH}, \eta) \tag{2.13}$$

This *defect-centric model* provides a very interesting relationship between the first two moments of the distribution and the corresponding *physical* parameters:

$$\eta = \frac{\sigma^2_{\Delta V_{TH}}}{2\langle \Delta V_{TH} \rangle}\ , \tag{2.14}$$

and

$$N_T = \frac{2\langle \Delta V_{TH} \rangle^2}{\sigma_{\Delta V_{TH}}}\ . \tag{2.15}$$

These observations lead to a *shift in the paradigm of reliability*. Even if the deterministic time-to-failure obtained from degradation extrapolation would be a realistic case, this number has to be replaced by a statistical distribution

for scaled devices. This means that although the *average* reliability of a gate stack might be sufficient for a certain operating condition, a *fraction* of the device population can still fail earlier. Knowing that the typical total population in BTI critical chip-elements such as an SRAM memory consists of ~millions of devices, less than 1 transistor per million (ppm) is allowed to fail in order for this element to function properly. In Fig. 42, the tolerable overdrive voltage distributions at continuous operation plotted for devices with identical gate dielectric properties but scaled area. Even though for all distributions the mean tolerable operation voltage is identical, the smallest devices show the largest fraction of failures, only reaching the targeted 1ppm at an overdrive voltage of a mere 0.1V

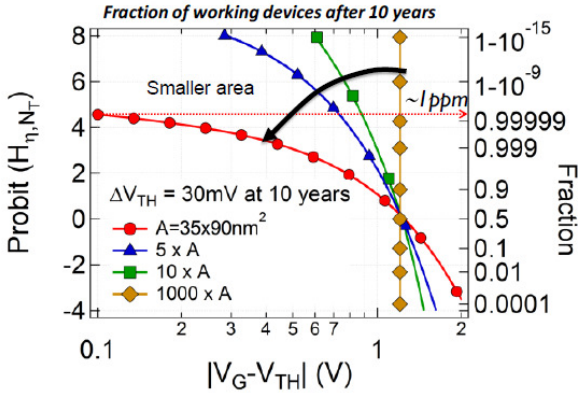



**Fig. 42: Tolerable overdrive voltage distributions at continuous operation plotted for devices with identical gate dielectric properties but scaled area. Even though for all distributions the mean tolerable operation voltage is identical, the smallest devices show the largest fraction of failures, only reaching the targeted 1ppm at an overdrive voltage of a mere 0.1V [replotted from Toledano11].**

## *2.2.10 Interpretations of the $V_{TH}$ shift*

After we have discussed the modeling efforts for BTI reliability in large-area and nanoscale devices, we will discuss in this Section the possible physical mechanisms that are responsible for these charge trapping events.

Up to date, it is still under discussion if the two invoked components in the *universal* model that were used to describe the degradation in large-area devices, i.e. the permanent and the recoverable component, are originating from the same microscopic processes or if they are an ordinary consequence of the wide spectrum of the trapping centers kinetics.

As discussed in the first Section of this chapter, two kinds of traps can capture charges: interface traps and oxide traps. The interface traps can accept charges from both silicon conduction and valence bands, as they are located at the $Si/SiO_2$ interface within the silicon bandgap. The occupancy of these states is understood to depend on the alignment of the Fermi level when the device is in equilibrium. The origin of these states arises due to the presence of dangling bonds at the interface. The existence of these dangling bonds is due to the natural mismatch of the lattice of Si and the amorphous layer of $SiO_2$.

### *2.2.11 The capacitance-voltage technique*

A possible techniques to monitor these interface states is with the capacitance-voltage (*C-V*) technique. The interface states, illustrated in Fig. 43, will acquire a net charge (either positive to neutral or neutral to positive depending on their type) during the sweep which will result in an overall stretch-out of the *C-V* sweep in the depletion region. Indeed, the interface states will act as a parallel capacitance to the depletion capacitance (at low frequencies). In a similar way, these interface states will result in a stretch-out of the *I-V* sweep below threshold, i.e. as a degradation of the subthreshold swing (SS). Finally, due to remote scattering mechanisms, these interface states will cause a reduction of the carrier mobility in the linear and the saturation regimes.

The other place where charge can be captured is in the bulk of the oxide, and is nominated $Q_{OT}$. These charges are associated with defects in the oxide layer, or at the interface between various oxide layers, but distant from the semiconductor interface. In contrast to interface charges, oxide states have much larger time constants for trapping and de-trapping. The trapping and de-trapping of charges in these defects is only indirectly visible by electrostatic screening (i.e. $V_{TH}$ shifts) during dynamic gate sweeps, shown in Fig. 44. Depending on the constitution of the gate dielectric, either the interface

charges or the oxide charges can be the dominant factor for the total $V_{TH}$ shift. It was shown by Aoulaiche, based on the charge-sheet-approximation, that for high-k stacks the dominant contribution of the $V_{TH}$ shift is the charging of oxide traps [Aoulaiche05].

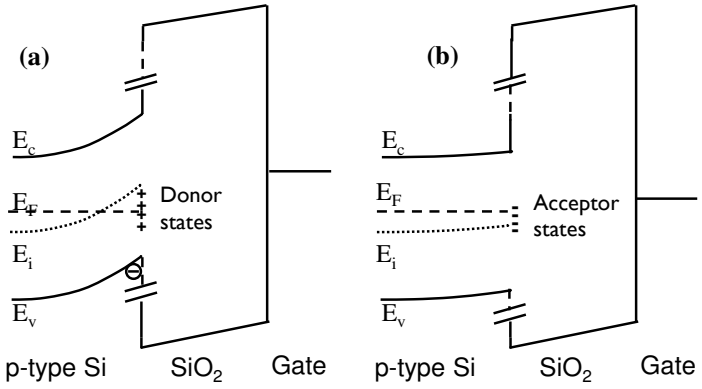



**Fig. 43: Typical Si/$SiO_2$ interface consisting of donor and acceptor states which are charged (a) positive to neutral and (b) neutral to negative respectively depending on the alignment of the Fermi level.**

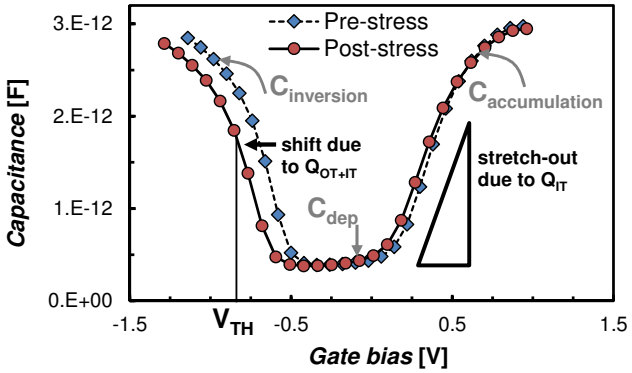



**Fig. 44: Pre- and post-stress *C-V* shows the impact of both oxide as interface charge trapping on the characteristics, resulting in a stretch-out due to $Q_{IT}$ and a shift of $V_{TH}$ due to $Q_{IT}$ and $Q_{OT}$.**

An advantage of the large time constants of these oxide defects is that each defect impact $\eta$, exponentially distributed as described in Section 2.2.9, can be measured. It is possible to identify each oxide defect's individual signature, which allows for example to trace its properties under various stress conditions and temperatures. This kind of testing methodology is called time-dependent-defect-spectroscopy (TDDS) [Grasser10b].

## 2.2.12 Physical origin of BTI – Reaction-diffusion model

In the 70's a physical model based on the breaking or 'de-passivation' of Si-H bonds, forming dangling silicon bond defects and the diffusion of the charged H-related species into the oxide was developed by Jeppson [Jeppson77]. Further work and modifications on this so-called 'reaction-diffusion' (R-D) model have been performed in the early 00's [Alam03, Mahapatra05]. The model has its origin in the fact that the observed degradation, is directly related to the surface-trap density, and was found to be proportional with $t_{stress}^{1/4}$, consistent with what one would expect from a diffusion-limited process (Fig. 45(a)). Moreover, it has been found that indeed a hydrogen passivation treatment, such as forming gas anneals or a deuterium-treatment, which forms stronger bonds with silicon than hydrogen, can severely impact the device's $V_{TH}$ instability, depicted in Fig. 45(b) [Schroder08, Pantelides07].

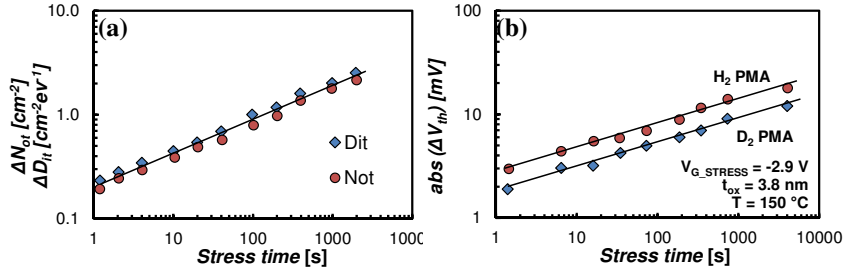



**Fig. 45: (a) The strong correlation of oxide trap and interface trap formation during BTI stress and (b) the positive impact of deuterium (D) anneals on BTI-induced threshold voltage shifts, both support the R-D model, suggesting that a hydrogen-related reaction should play a central role during the degradation [replotted from Schroder05].**

In the last decade however, mostly with the introduction of fast(er) measurement techniques, this R-D model was found to exhibit *several inconsistencies with experimental findings*. Even though it is not the purpose of this work to give an exhaustive explanation of the deficiencies of the R-D model, some of these findings can help to get an idea of the physical mechanisms behind the BTI degradation in MOSFETs.

To give an example, the R-D model, in which the hydrogen diffusion is believed to be the time-limiting step following Fick's law of diffusion, predicts a "universal recovery", i.e. a strong link between the stress time of the oxide and the relaxation time. Approximately, the normalized RD recovery can be written as [Grasser07a]:

$$\frac{\Delta V_{TH}(t_{stress}, t_{relax})}{\Delta V_{TH}(t_{stress}, 0)} = \frac{1}{1+(t_{relax}/t_{stress})^{1/2}}. \quad (2.16)$$

Indeed, this universal recovery will appear as a logarithmic relaxation for a few decades, as experimentally observed, but the overall relaxation is only expected to occur for about 4 orders of magnitude in time. However, experimentally the recovery is already vigorously appearing from the microsecond timescale, independent of $t_{stress}$ and continuing over at least 10 decades of time, as shown in Fig. 46(a), where the threshold voltage shift relaxation after 100ks of stress begins in instantly, which is not according to the predicted R-D relaxation in Eq 2.16.

Moreover, the relaxation is expected to decrease with stress time, as the recovery is explained by re-passivation by leftover hydrogen-related species near the interface. Experimentally the contrary was shown however, as depicted in Fig. 46(b): the amount of interface state recovery per cycle, expressed as variation in charge-pumping current, $\Delta I_{CP}$, is observed experimentally to remain unchanged after many stress and recovery cycles whereas the R-D model predicts a steady decrease of interfacial trap recovery because the hydrogen is transported further away from the interface.

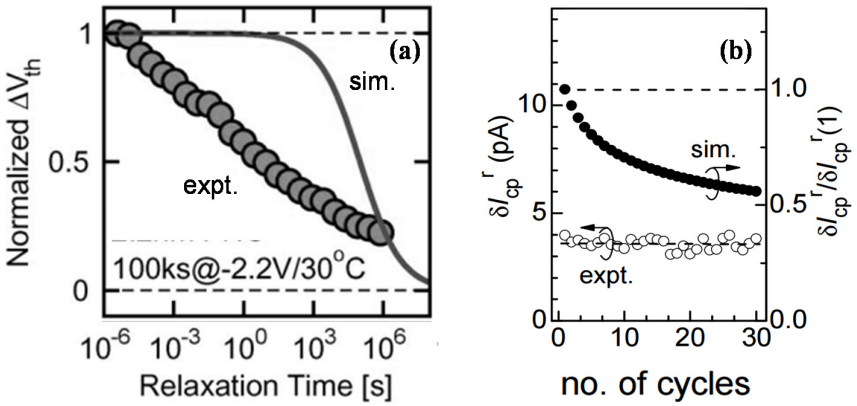



**Fig. 46: (a) The threshold voltage shift relaxation after 100ks of stress begins in instantly, which is not according to the predicted R-D relaxation in Eq 2.16 and (b) amount of interface state recovery per cycle, expressed as variation in charge-pumping current, $\Delta I_{cp}$, is observed experimentally to remain unchanged after many stress and recovery cycles whereas the R-D model predicts a steady decrease of interfacial trap recovery because the hydrogen is transported further away from the interface [replotted from Grasser11 and Teo09].**

Both effects were dealt with by somewhat artificial modifications of the R-D model: the $H_2$ diffusion was modeled as being dispersive, i.e. being time-dependent [Kaczer05]. Also a hole-trapping component was added, active in the sub 1 second region of the relaxation, and subsequently counterbalanced by the long-term $H_2$ diffusion [Mahapatra09]. It remains however doubtful that the experimentally observed relaxation on large time-scales always shows a continuous "log $t_{relax}$" relaxation for large devices, and never showed a characteristic 'bump', as would be expected if the recovery process originated from two distinctly different processes [Grasser11].

Concluding, we can state that the main implication from the proposed R-D models is that the observed relaxation time constants must be somehow inherent to the charge trapping itself, rather than being a consequence of a diffusion process.

## 2.3 Random-telegraph-noise

Random-telegraph-noise (RTN) is another important dynamic variation source in ultra-scaled MOSFETs and can be strongly linked to BTI. A typical observation of RTN is that the current is fluctuating randomly between several discrete stages within a broad range of timescales. RTN can be seen as the discrete small-scale equivalent of low frequency *1/f*-noise observed in large area devices.

In the past, many publications have shown that there is a strong link between noise-related phenomena and BTI [Kaczer08, Kaczer10, Grasser09b, Ang08]. Moreover, it was shown that BTI and RTN-defects share a large portion of their properties, such as their occurrence distribution $P_N$, their impact distribution $\eta$ et cetera [Grasser09b]. Capture and emission times of RTN defects are typically also voltage dependent like BTI defects, and therefore most of the RTN defects will also contribute to BTI recovery. Finally, both BTI and RTN defects are volatile, meaning that they can disappear and reappear over time.

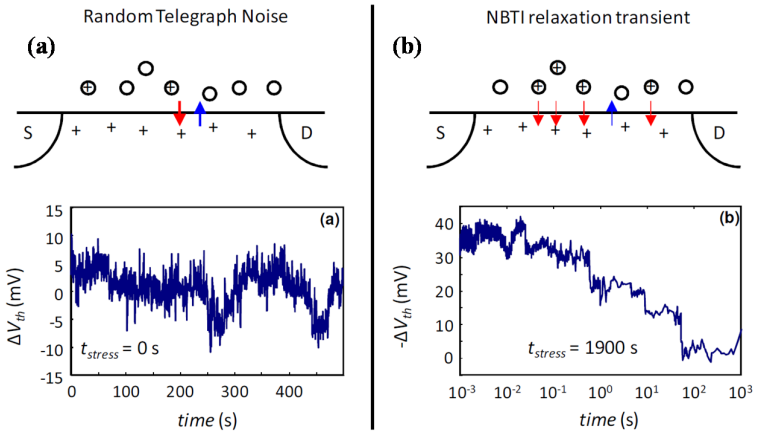



**Fig. 47: (a) Dynamic equilibrium of charging and discharging of oxide defects at a constant gate bias, giving rise to fluctuations in the channel current. (b) After BTI stress, the system converging towards its dynamic equilibrium by discharging an excess amount of oxide defects. [replotted from Kaczer11]**

RTN traps exhibit however one "additional" property: at certain measurement conditions, the defect's capture and emission probabilities and corresponding capture and emission times are similar. At these conditions, an equilibrium of charging and discharging these oxide defects will result. If the device is stressed before, a net discharge of the excess trapped charges will result subsequently in BTI relaxation, until the device returns into the original equilibrium. Both mechanisms are illustrated in Fig. 47.

## 2.4 Channel-Hot-Carrier degradation

Hot carriers (HC) are carriers that attain a very high kinetic energy from being accelerated by a high electric field in short-channel MOSFETs. These highly-energetic particles can collide with the lattice, and generate electron-hole pairs by impact ionization, or can get injected into the gate dielectric, where they can become trapped or cause interface states to be generated. They can also reach the gate and contribute to gate leakage current.

Hot carrier generation, injection and the *resulting device degradation* have become one of the major reliability issues in current MOSFET technology. Mainly due to the ever-decreasing MOSFET channel lengths, the lateral fields are increasing towards the ~1 MV/cm range.

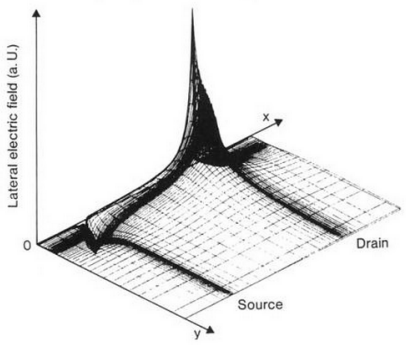



**Fig. 48: Lateral electric field of a conventional transistor operating in saturation as given by drift-diffusion calculations. The peak electric field near the drain region will allow certain carriers to become "hot" [replotted from Weber88].**

Typical conditions for the generation of these hot carriers are when the $V_{DD}$ (i.e. the operating voltage of the chip) is applied on both the gate as the drain terminal of the transistor. This will result in a peak in the lateral field in the *pinch-off* region, near the drain junction, illustrated in Fig. 48. It is this electric field peak that will enhance the hot carrier generation.

Reducing this lateral electric field peak will thus help to reduce hot carrier degradation. An engineering solution applied already in the '90's, and still used up to now, is the so-called lightly-doped-drain (LDD) [Ogura80] or the graded channel (GC) structure [Lyu97]. There are, however, many innovations in recent CMOS scaling, typically to prevent short-channel effects (SCE) such as the introduction of HALO implants to prevent punch-through, that *enhance* the electric field and thus also the hot carrier degradation.

In the Section 2.4.1, a qualitative description of the injection of holes and electrons under the influence of external voltages is discussed, in order to emphasize the conditions where holes and electrons could play a role in the degradation mechanisms. A brief overview of the experimental techniques will be given. Section 2.4.3 will discuss some experimental results obtained under uniform and non-uniform (i.e. the real operating conditions) hot-carrier injection conditions for planar and multi-gate devices.

### *2.4.1 Phenomenological overview*

For the injection of hot carriers into the dielectric there are four distinctive injection mechanisms [Takeda83]: channel hot-electron (CHE) injection, drain avalanche hot-carrier (DAHC) injection, secondary generated hot-electron (SGHE) injection, and substrate hot-electron (SHE) injection:

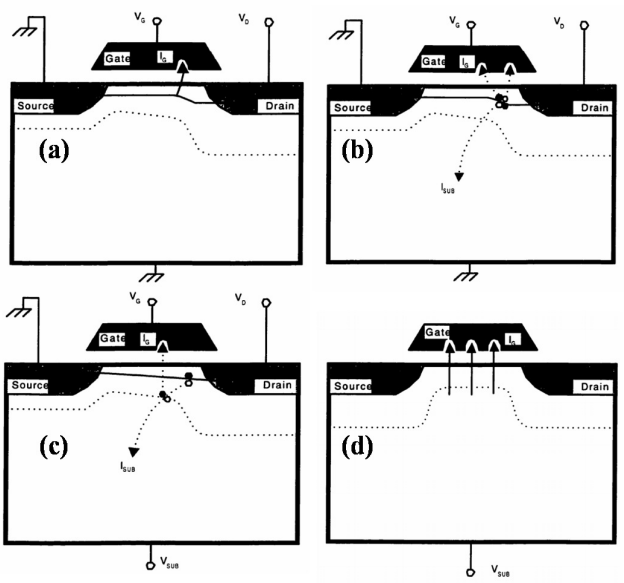



**Fig. 49: Schematic illustrations of hot carrier generation mechanisms: (a) Channel hot carriers (CHC), (b) Drain Avalanche Hot carriers, (c) Secondarily Generated Hot Carriers (SGHC); an d(d) Substrate Hot Carriers (SHC), replotted from [Liu05]**

As schematically shown in Fig. 49(a), carriers flowing from source to drain, without undergoing any collision that causes energy loss (so called "lucky electrons") might acquire enough energy to get injected into the gate oxide.

Drain Avalanche Hot Carriers (DAHC) are due to the high electric field near the drain region promoting avalanche multiplication, depicted in Fig. 49 (b). If $V_D$ is large, a reduction of $V_G$ increases the electric field at the drain to the point where avalanche multiplication due to impact ionization may substantially increase the supply of both hot electrons and hot holes. These are subsequently injected into the gate in similar as channel hot electrons. This degradation mechanism is considered to be the most severe degradation at room temperature.

Secondary generated hot-electron (SGHE) injection: as shown in Fig. 49 (c), when the electric field is very high as described in the DAHC conditions,

the newly-generated hot carriers are able to cause *secondary impact ionization* in the depletion region during their journey to the substrate. The electrons generated due to secondary impact ionization can also be injected into the gate and cause additional degradation. This secondary effect only becomes a problem in nanoscale devices.

Substrate hot-carrier (SHC) injection (Fig. 49(d)): in this case, hot carriers are thermally generated or injected by means of an external forward-biased pn-junction injection structure. The SHC injection occurs when a large substrate bias is applied. In this case, the channel carriers will gain energy from the high electric field in the surface depletion region. In contrast to the other described degradation mechanisms which will cause localized degradation, this mechanism will cause an *uniform degradation* along the interface of the channel. This mechanism is well-suited for accelerated testing or an in-depth understanding of the phenomenon as multiple stress mechanisms can be controlled separately: the oxide field is controlled by the gate voltage, the amount of carriers is controlled by the bias of the external carrier injecting diode, and the energy is controlled by the substrate bias.

Typically hot carrier degradation is monitored in nMOS devices, since they are expected to behave worse than their pMOS counterparts, due to multiple reasons:

- the carrier mobility, which is dependent on the mean free path and thus related to the chance of having “lucky electrons”, is higher than in pMOS devices,
- the barrier for holes at the $Si/SiO_2$ interface is higher,
- and the efficiency of electrons for generating electron-hole pairs is higher.

Although a lot of understanding has been gained in the past decades, there still remains some controversy and debate about the nature of degradation and the involved physical processes. This is, to a large extent, due to the limitations of the conventional techniques used to monitor carrier injection and device degradation. Hot-carrier injection and degradation are very localized phenomena and the extraction of correct and unambiguous information about the nature and magnitude of the resulting *non-uniform defect distribution* is not straightforward and often not even possible.

### 2.4.2 Basic interpretation of hot carrier generation

There is no consensus on the physical processes involved in the generation and degradation process due to the still-lacking precise understanding. Yet it is believed that the present HCI models are sufficiently reliable and universal for present transistor features.

The Lucky Electron Model, as described by [Hu85] is up to this date firmly established as the *guiding principle* of most hot carrier models and lifetime projection methodologies. However, many recent observations are no longer corresponding with this model, and modifications had to be made as will be discussed further in this Section. The model is based on the fact that the probability of a channel carrier to be accelerated to a given energy $E$ has to be described by an exponential distribution:

$$f(E) = P(E) = e^{-E/q\lambda F} \tag{2.17}$$

depending on the carrier's mean free path $\lambda$ (typically up to 10nm for carriers in Si at room temperature), $q$ the elementary charge and $F$ the electric field. Interestingly, as the mean free path is known to be reduced due to enhanced electron-phonon scattering at elevated temperature, the amount of 'lucky electrons' and thus hot carrier generation is expected to *reduce* with temperature. This is in contrast to BTI where degradation was shown to be accelerated by T and to follow an Arrhenius law.

To model the impact ionization rate, it is assumed that this rate is determined by carriers having an energy above a certain energy threshold $\phi_i$. The drain current acts as a source (the supply of carriers) and the peak field is used together with the (hot-carrier) mean free path to describe the ionization probability:

$$\frac{I_{SUB}}{I_D} = Ae^{-\phi_i/q\lambda F_m} \ . \tag{2.18}$$

with the ratio of substrate current $I_{SUB}$ to drain current $I_D$ being the impact ionization ratio or 'multiplication factor M', $F_M$ the maximum field in the channel. Carriers with an energy below $\phi_i$ are supposed to yield no damage. Therefore, the multiplication factor can be used as an indirect measurement of the electric field that is generating hot carriers, using Eq. 2.16.

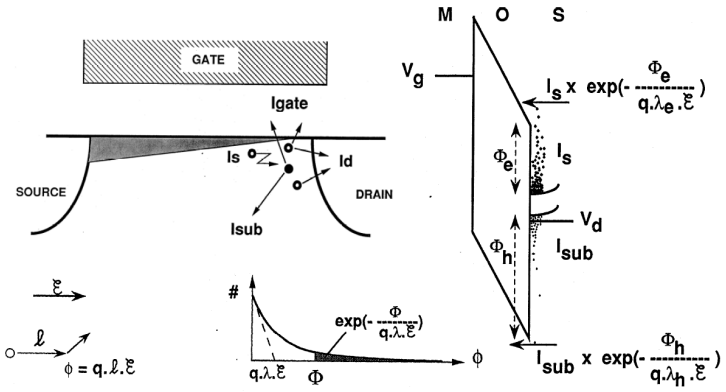



**Fig. 50: Illustration of HC impact ionization at the drain side of a MOSFET operated in $|V_{DS}| > |V_{GS}| - |V_{TH}|$. According to the LEM, only the carriers that obtained sufficient energy (i.e. more than the semiconductor's bandgap) can generate electron-hole pairs by impact ionization. The highest energetic carriers can directly overcome the barriers between the semiconductor and the insulating layer due to tunneling phenomena [replotted from Sze81].**

Similarly to BTI induced degradation, the injection of hot carriers into the gate dielectric of the MOSFET causes oxide trapping and interface state and oxide defect generation. This has an effect on the *I-V* characteristics of the device, i.e. a shift of the threshold voltage, subthreshold slope degradation and reduction of the device transconductance. In the derivation below, we will estimate the degradation and the lifetime by means of a critical interface state generation $\Delta N_{IT}$.

### *2.4.3 Lifetime prediction techniques and observations in multi-gate devices*

Hot carrier degradation is one of the most complex degradation phenomena to make lifetime predictions, as it always comes convoluted with of a mix of other degradation phenomena. As a result, there is always also a BTI-induced charge trapping effect, the temperature state of the transistor in DC bias that might strongly differ from that during AC bias (i.e. the "self-heating" effect),

et cetera. All these effects should be encompassed in the modeling in order to make relevant predictions for device lifetime operation.

Typical parameter shifts such as $\Delta N_{IT}$, $\Delta SS$, $I_{DLIN}$ or $I_{DSAT}$ reduction can be used as degradation monitors. It is however important to note that for hot carriers degradation in typical operating condition, i.e. due to a lateral field in the channel, the degradation is localized at the drain side where the electric field peaks. This localized degradation will induce an asymmetric impact on the device's *I-V* characteristic, depending on the sign of the applied $V_{DS}$.

The typical experimental procedure is as follows: an initial $I_D$-$V_G$ characteristic is measured on a pristine device, a constant stress voltage is applied on gate and drain with source and bulk grounded. At specific intervals the stress is interrupted and a full $I_D$-$V_G$ characteristic is measured to monitor the degradation of $V_{TH}$, $I_D$, subthreshold swing, and $g_m$ as a function of the stress time. Even though the MSM-technique, as discussed in Section 2.2 can also be used, it would neglect the $g_m$ or subthreshold slope information. It should be noted however that some of the recoverable part of the CHC degradation can be lost during the full $I_D$-$V_G$ measurement. Intermittent $I_D$-$V_G$'s during CHC are depicted in Fig. 51(a), and the extracted $V_{TH}$ shift is depicted in Fig. 51(b).

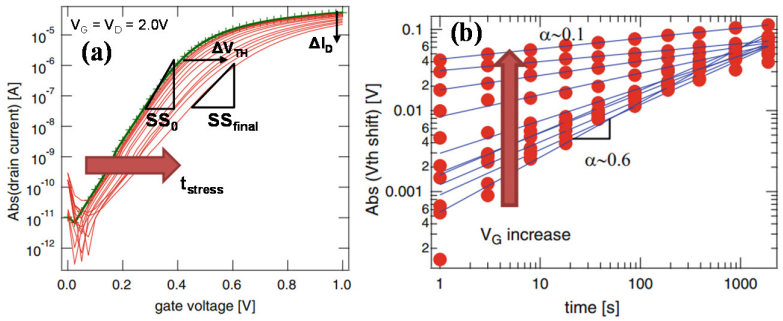



**Fig. 51: (a) Illustration of $V_{TH}$ shifts and as a function of stress time in long channel n-FinFETs. (b) The $V_{TH}$ shift is extracted from $I_D$-$V_G$ curves measured in saturation-reverse mode. The time exponent decreases at higher $V_{G\_STRESS}$ [replotted from Cho14].**

Based on the observed time exponents of the HC degradation, conclusions can be drawn on the nature of the degradation: if the interface degradation by

hot carriers is dominant, a time power-law exponent between 0.5 and 1.0 should be observed [Hu85]. Lower power law time exponents—between 0.1 and 0.2—are expected when the carrier trapping from the substrate into the oxide bulk defects is the main degradation mechanism. This is illustrated in Fig. 52.

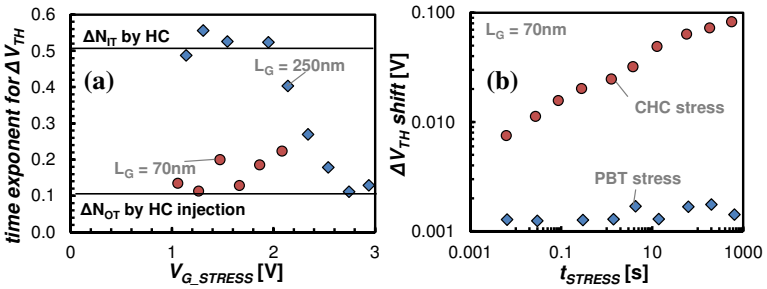



**Fig. 52**: **(a) Extracted time exponents *n* from Fig. 51. The *n* of long $L_G$ devices decreases at higher $V_G$, indicating a change in degradation mechanism from interface dominated towards hot/cold carrier injection regime. For short channel carrier injection is dominant in all stress regimes (b) CHC vs PBT stress for short channel devices, indicating negligible charge injection from 'cold' carriers [replotted from Cho14].**

For lower $V_{G\_STRESS}$, the time exponent *n* saturates around 0.5, indicating the interface degradation from the hot carriers is dominant. However, when $V_{G\_STRESS}$ increases, the time exponent decreases and saturates at a value of ~0.1. The saturation of the time exponent is typically seen in PBTI, i.e. in cases where the carrier injection into the bulk oxide defects is dominant. Fig. 52 additionally shows the time exponent on short channel nFinFETs, being continuously lower than 0.2. This implies the hot carrier injection is dominant in all stress regimes.

Finally, to *distinguish between hot and cold* carriers affecting the CHC reliability in short channel devices, PBTI measurements can be performed separately from the CHC measurement, i.e. by applying low $V_D$ during the stress. As can be seen from Fig. 52(b), while the CHC stress induces a clear $\Delta V_{TH}$, the PBT stress barely degrades the device. At the PBT stress condition, only the 'cold' carriers, are reacting with the gate oxide to generate device degradation. In the HC stress condition, impact ionization can occur as the

lateral electric field is high, and therefore both hot as cold impact the overall device degradation.

## 2.5 Stress induced leakage and oxide breakdown

Similar to bias temperature instabilities and hot carrier degradation, (stress-induced) oxide leakage and breakdown also occurs as a result of high gate voltage stress. However, in this case the stress conditions are *generating* electronically active defects until an irreversible breakdown or '*percolation*' path is formed between the channel and the gate, illustrated in Fig. 53.

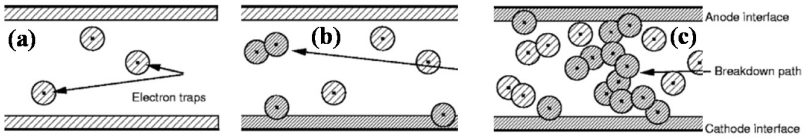



**Fig. 53: Different stages during oxide stress: (a) pre-existing neutral traps in the oxide, (b) formation of conductive paths due to defect clustering (c) leading to the creation of a conductive breakdown path from anode to the cathode [replotted from Hull05].**

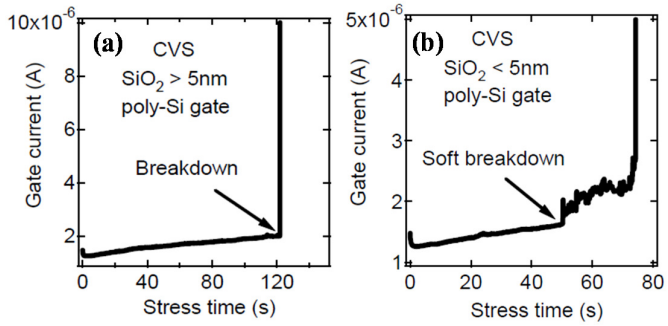



**Fig. 54: Illustration of (a) hard breakdown under constant-voltage stress, typically observed in devices with EOT > 5nm, (b) whereas in devices with EOT < 5nm, current noise fluctuations are observed before the final hard breakdown.**

Typically oxide breakdown is observed as a gradual or abrupt increase in the gate leakage current during a constant gate voltage stress (Fig. 54). The

abrupt, *breakdown* event itself can be characterized as *hard* if a short (i.e. an Ohmic resistive path) between the gate and a junction is occurring, whereas a *soft breakdown* preserves the normal operation of the transistor, albeit with increased gate leakage current.

### 2.5.1 Physical origin of dielectric breakdown

*Extrinsic* TDDB occurs during the early life of the device and is related to the presence of a weak spot or defects, for example due to non-uniform oxide deposition. These breakdowns are thus the result of poor processing and shall not be considered here.

*Intrinsic* TDDB however, can occur even uniform and defect-free oxides during the device operation due to an electrical stress-induced generation of oxide traps. The mechanism leading to intrinsic breakdown is generally understood as follows:

- charge injection into the oxide at high oxide fields generates oxide traps,
- additional *trap-assisted leakage paths* due to these oxide traps cause a gradual increase of the gate current. The increase is typically denominated as stress induced leakage current or SILC,
- formation of a *conduction path* from the cathode (substrate) towards the anode (gate) if enough oxide traps are formed nearby each other, visible as an abrupt increase of the gate leakage current.

In contrast to BTI and HCI degradation described in the above Sections, the physical mechanism for TDDB must thus incorporate a defect generation step. Typical origins are thought to be trap creation or hydrogen release [Grasser14], anode hole injection and thermo-chemical electric field [Hull05].

## 2.6 Self-heating effect

It is widely known that changes in temperature of a semiconductor chip will affect the system's speed, power consumption and reliability. One of the most cited roadblocks in semiconductor scaling is the power problem, i.e. power

densities, heat generation and chip temperatures reaching levels that can jeopardize the reliable operation of the integrated circuit.

The issue of *global* chip heating is compounded with the transistors' *local* self-heating effect (SHE), illustrated in Fig. 55. This local self-heating effect can be described as nanoscale hotspots caused by the extremely high local current densities. The temperature effects related to device self-heating do not only impact the transistor's performance ad hoc, but will also cause associated effects that impact device and interconnect reliability.

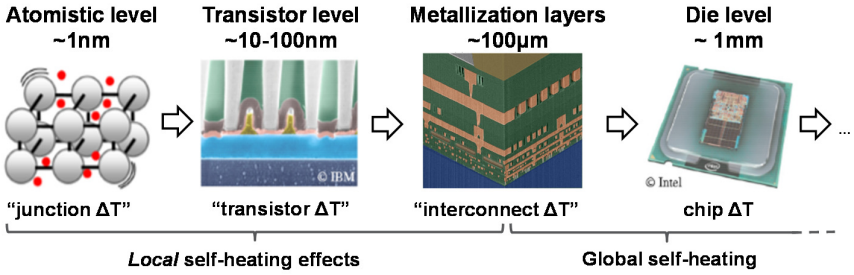



**Fig. 55: Illustration of scales at which the heating effects can be treated in a typical semiconductor system. The *actual* "local" operating temperature of the device during operation will depend on the series-chain of boundary conditions set by the scales upward in the hierarchy.**

In this Section, we discuss the main origins of device self-heating and explain the physics behind this effect. We also discuss the existing modeling and measurement methods.

### 2.6.1 Main origin of self-heating

As the transistors are powered up and drive currents in circuit operation, Joule heating will occur due to resistive losses. This power dissipation will subsequently lead to a (local) temperature increase, which can be strongly non-uniform and cause local hot-spots, depending on the heat conductivity of the system. The operating temperature of a device $T_{device}$ can be written as:

$$T_{device} = T_{ambient} + \Delta T_{chip} + \Delta T_{SHE} \quad (2.19)$$

with $T_{ambient}$ the ambient temperature, $\Delta T_{chip}$ the temperature increase of the chip during operation and $\Delta T_{SHE}$ the temperature increase in the channel of the transistor.

The steady-state $\Delta T_{SHE}$ is assumed to be proportional to the dissipated power $Q_{dissipated}$ and the thermal resistance $R_{TH}$ between the channel and the external boundaries of the system, i.e.

$$\Delta T_{SHE} = Q_{device}.R_{TH} = I_D.V_D.R_{TH} \quad . \quad (2.20)$$

The major challenge associated with SHE assessment is the evaluation of this $R_{TH}$. Since MOSFETs shrink down continuously, $R_{TH}$ is expected to increase because a reduced silicon volume is available for heat removal. The main reason for this is the discrepancy (a factor of 100 difference) in heat conductivity of silicon, i.e. the active area of the transistor and the shallow trench isolation (STI) around it, typically consisting of $SiO_2$. Therefore, a lot of learning about self-heating has already been gained in the last decade in Silicon-On-Isolator (SOI) devices, in which the heat transfer towards the bulk is impeded by the buried $SiO_2$ underneath the channel [Dallman95, Fiegna08].

In addition to that, $V_{DD}$ is not being scaled down accordingly below 65nm node [Groeseneken10], and the power (density) $Q_{device}$ in the channel is expected to increase.

Novel and sometimes complicated device geometries tend to make heat removal from the channel region more difficult, illustrated in Fig. 56. Therefore, SHE is thus expected to become more pronounced in the upcoming device nodes. Moreover, most new materials introduced in device processing also exhibit lower thermal conductivities than bulk silicon, as illustrated in Table II.

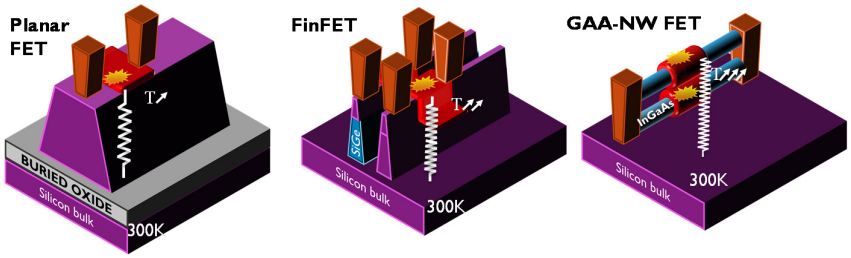



**Fig. 56: Evolution of transistor designs from bulk and SOI planar devices, towards transport-enhanced (by thin body and/or strain) fully-depleted FinFET devices up to the ultimate CMOS device: the gate-all-around nanowire incorporating high-mobility channel materials. Devices with narrow bodies exhibit poorer thermal properties mostly due to the omnipresent STI (oxide) surroundings.**

**Table II: Thermal conductivities ($\kappa_{TH}$) of a several bulk materials typically used in MOSFETs. Note the significantly lower $\kappa_{TH}$ of thin film silicon w.r.t. bulk material due to phonon boundary scattering.**

| Material | Occurrence in system | $\kappa_{TH}$ @ 300K [$W.K^{-1}.m^{-1}$] |
|---|---|---|
| Si | Substrate | 148 |
| | thin film fins (<20nm) | 20-10 |
| $SiO_2$ | STI | 1.38 |
| NiSi | S/D contacts | 16 |
| W | Local interconnects, VIA's | 40 |
| Cu | Metal layers | 250 |
| $HfO_2$ | Gate dielectric | 0.5-1 |
| TiN | Gate liner | 11 |
| TaN | Gate liner | 3 |
| Al | Gate metal | 40 |
| GaAs | Fin buffer | 46 |
| InGaAs | High-u channel (nMOS) | 5 |
| $SiGe_{50}$ | High-u channel (pMOS) | 11 |
| InP | Fin buffer | 68 |
| InAs | High-u channel (nMOS) | 25 |
| Ge | High-u channel (pMOS) | 60 |

Finally, at nanometer scales the properties of these materials are modified. Indeed, as devices are scaled to dimensions comparable to, or less than the mean free path of the thermal energy carriers (i.e. the phonons), the materials will no longer behave as bulk materials, but additional scattering effects will come into play.

On large-scale devices, the steady state heat transfer can be treated by the continuum model, i.e. Fourier's law, which is expressed intrinsically as:

$$\vec{q} = \kappa_{TH} \nabla T \tag{2.21}$$

with $q$ being the heat flux density, $\kappa_{TH}$ the material's thermal conductivity and $T$ the temperature, and in the extrinsic form as:

$$\Delta T = R_{TH} Q \tag{2.22}$$

with $R_{TH}$ the thermal resistance and $Q$ the total heat flux. The time-dependent model is then described as:

$$C_S \frac{\delta T}{\delta t} = \nabla \cdot (\kappa_{TH} \nabla T) + \vec{q} \tag{2.23}$$

where $C_S$ is the heat capacity per unit volume.

For *metals*, the thermal conductivity $\kappa_{TH}$ is directly related to the electrical conductivity $\sigma$ through the empirical Wiedemann-Franz law (later refined by Lorenz), as the particles responsible for heat transport are the nearly-free conduction electrons:

$$\kappa_{TH} = \frac{\pi^2}{3} \left( \frac{k_B}{q} \right)^3 \sigma T \tag{2.24}$$

with $k_B$ the Boltzmann constant and q the elementary charge [Wiedemann53].

The energy carriers responsible for heat transport in *dielectrics* and semiconductors are the lattice vibrations or phonons. For these materials, the bulk thermal conductivity can be described as:

$$\kappa_{TH} = \frac{1}{3} C_S \langle v \rangle \Lambda \tag{2.25}$$

where $v$ is the phonon group velocity and $\Lambda$ the average phonon mean free path [Ziman60].

Equation 2.25 has an impact on nanoscale systems: not only is the heat *transfer* governed by phonon and electron transport, the heat *generation* in nanoscale devices is originating from energy transfer between conduction carriers and the medium, as is illustrated in Fig. 57. The primary path of energy transport is represented first by scattering between electrons and optical phonons and then optical phonons to the lattice [Majumdar95]. Physically speaking, with the application of a drain bias, electrons gain energy from the lateral electric field and interact with both the optical and the acoustic phonon population. Since the optical phonons involved in the process have much larger energy than the acoustic phonons, most of the electron energy is transferred as heat in the optical phonon population. However, the optical phonons have nearly zero group velocity, i.e. they do not contribute to heat *transfer*, so the heat remains localized. The transfer of heat to acoustic phonons via anharmonic decay processes is relatively slow [Vasileska14].

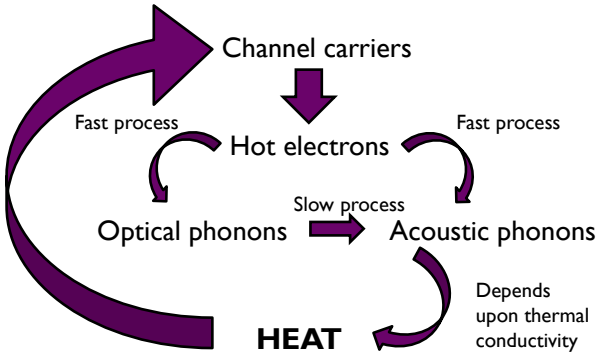



**Fig. 57: Illustration of heat generation and transport in nanoscale devices. The conduction carriers being accelerated by a high electric field in the channel become hot and can scatter on the lattice ions, resulting in optical and acoustic phonon generation. Due to their low group velocity, optical phonons are do no *conduct* heat, in contrast to acoustic phonons.**

When simulating nanoscale or "confined-geometry" devices, such as FinFET or nanowire devices, the *granularity* of heat transport via phonons must be considered. For example, the intrinsic thermal properties of materials

at nanometer length scales are modified, due to enhanced phonon-boundary scattering. Fig. 58 illustrates the heat transport in a lattice by phonons, in which the phonon mean-free-path can be reduced by phonon-impurity, phonon-phonon and phonon-boundary scattering.

Also the temperature dependency of thermal conductivity, due to phonon-phonon scattering should be taken into account. Various methods to pragmatically take into account this granularity in device simulators will be further elaborated in Chapters 5 with electro-thermal simulations and in Chapter 6 with the use of temperature dependent conductivity *tensors*.


**Fig. 58: Illustration of heat transport in a lattice by phonons, in which the phonon mean-free-path can be reduced by phonon-impurity, phonon-phonon and phonon-boundary scattering.**

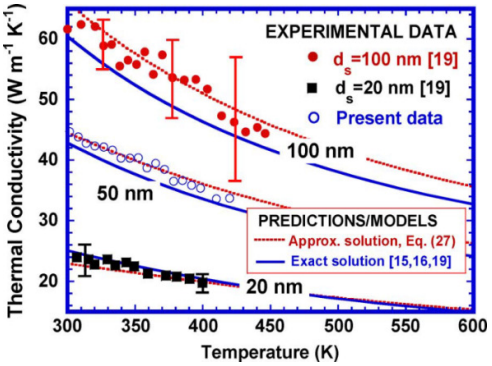



**Fig. 59: Measurements and simulations of thickness and temperature dependence of the silicon thermal conductivity caused by phonon-phonon and phonon-boundary scattering mechanisms. [replotted from Liu06]**

### 2.6.2 Temperature effects on device performance

The drive current of the transistor depends on the $\mu$, $V_{TH}$ and $v_{sat}$. The temperature relationships for these parameters can be expressed using the following empirical equations [Morshed09]:

$$\mu(T) = \mu_0 \left(\frac{T}{T_0}\right)^{\alpha_\mu} \tag{2.26}$$

$$V_{TH}(T) = V_{TH0} + \alpha_{V_{TH}}(T - T_0) \tag{2.27}$$

$$v_{sat}(T) = v_{sat0} + \alpha_{v_{sat}}(T - T_0) \tag{2.28}$$

where $T$ is the temperature, $T_0$ the nominal temperature; $\mu_0$, $V_{TH0}$, and $v_{sat0}$ are the respective mobility, threshold voltage, and saturation velocity at $T_0$, and $\alpha_\mu$, $\alpha V_{TH}$, and $\alpha_{vsat}$ are empirical parameters describing the mobility, threshold voltage and saturation velocity temperature coefficients. In the concerned temperature region, $\mu$, $V_{TH}$, and $v_{sat}$ all *decrease* with increasing temperature; typical values for the coefficients are $\alpha_\mu$ ~ -1.3, $\alpha_{VTH}$ ~ -0.8 mV/K, and $\alpha_{vsat}$ ~ -97 m/(s·K) [Sze81, Filanovsky01].

For conventional long-channel MOSFETs, due to the large amount of scattering events in the channel, the saturation current is mainly determined by the mobility values, which is known as the "drift-diffusion" limit. For short channel devices, the channel length can be shorter than the carrier mean free path (MFP). In that case, quasi-ballistic transport takes place and the carrier velocity $v_{sat}$ in the channel at the location corresponding to the peak of the source-side potential barrier will determine the MOSFET saturation current. The drive current is then described as follows:

$$I_D(T) = v_{sat}(T).W.P_S.[V_{GS} - V_{TH}(T)]^{\alpha} \tag{2.29}$$

where $P_s$ is a technology-specific constant and $\alpha$ is a technology-specific exponent. The impact of drive current on temperature can be described as the sum of the impacts:

$$\left.\frac{dI_D}{dT}\right|_{tot} = \left.\frac{dI_D}{dT}\right|_{v_{sat}} + \left.\frac{dI_D}{dT}\right|_{V_{TH}} \tag{2.30}$$

in which the first term is negative, and the second term is typically positive. In typical CMOS technologies at $V_{DD}$ conditions, the magnitude of the former term is greater than the magnitude of the latter one. In other words, the drive current will reduce at elevated temperatures, which is referred to as the *normal temperature dependence.*

However, as $V_{GS}$ approaches $V_{TH}$, changes in $V_{TH}$ have a larger impact on $I_{D.}$ Therefore, at lower supply voltages the drive current *increases* with temperature, which is referred to as a the *reverse temperature dependence*. The gate bias at which $dI_D/dT = 0$ is the zero-temperature-coefficient voltage ($V_{ZTC}$).

The temperature dependence of the $I_{D,}$ together with the thermal inertia of the device caused by its thermal capacitance $C_{TH}$, is typically used in indirect self-heating measurements, as will be discussed in Section 2.6.5.

### *2.6.3 Self-heating simulation techniques*

Roughly, we can classify the self-heating simulation techniques into four levels, according to their exactness, complexity but also uncertainty of the simulation parameters. This is schematically illustrated in Fig. 60.

Finite element simulations and simulations by commercial TCAD software, rely Fourier's law of heat diffusion and solve linear heat equations (level 1-2 [Prasad13]). Solving the heat diffusion equation in its steady state (Eq. 2.21) or time-dependent (Eq. 2.23) form is not very computationally intensive. Therefore, these simulations have the advantage that the *entire physical structure* of the device can be reproduced in the 3D finite-element-simulator (FEM), including the back-end-of-line (BEOL) surrounding layers. The

importance of simulating the entire BEOL line will be illustrated in Chapter 5.

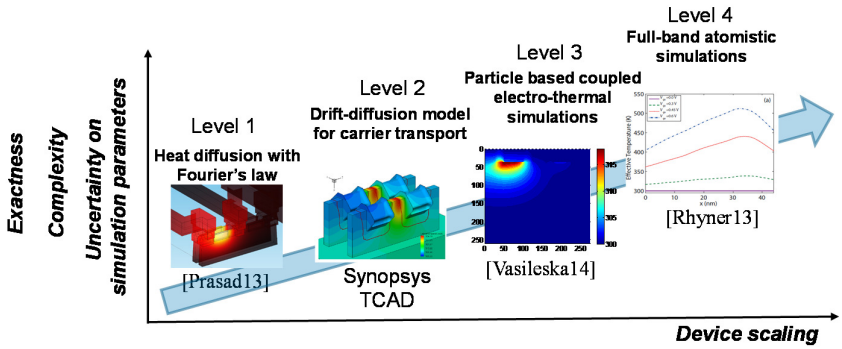



**Fig. 60: Illustration of various techniques to simulate the self-heating effect in nanoscale devices [Level 4 data replotted from Rhyner13].**

In typical TCAD simulations, the thermodynamic carrier transport model extends the standard drift-diffusion approach to account for electro-thermal effects, under the assumption that charge carriers are in thermal *equilibrium* with the lattice. In practice, it solves the lattice temperature (heat flow) equation iteratively in addition to the Poisson equation and carrier continuity equations [Synopsys14].

Other research groups have custom built simulation tools which rely on non-equilibrium statistical mechanics, and the simulation outcome is the result of a Monte-Carlo approach (level 3, [Vasileska14]). Briefly, in those more advanced academic simulators, the Boltzmann transport equation (BTE) for the electrons is self-consistently coupled to Poisson equation solvers and energy balance equations for the acoustic and optical phonons [Pop04, Vasileska10].

Practically, the electron energy and density distributions are calculated and subsequently the electron-lattice interactions are modeled. More accurately, the optical and acoustic phonon-electron and phonon-phonon interactions are modeled separately. In this case, the phonon distributions are calculated by using the energy balance equations. The 2D Poisson equation is then solved self-consistently with a Monte Carlo transport kernel and 2D energy balance

equation solvers for the acoustic (lattice) and optical phonon populations, as depicted in Fig. 61. These phonons will be eventually converted into lattice heat, which depends on lattice thermal conductivity.

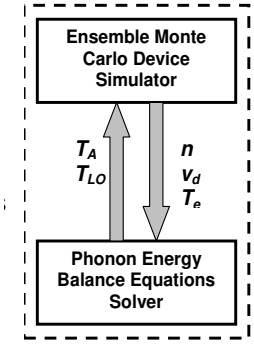



**Fig. 61. Practical implementation of the ensuring self-consistency in between the MC device simulator and the energy balance equation solver [replotted from Vasileska14].**

For details regarding the full-band atomistic simulations (level 4), solving the coupled Boltzmann transport equation *both* for electrons and phonons, we refer to [Rhyner13].

It should, however, be noted that the goal of this work is not to develop such a fully self-consistent simulator, as it has been demonstrated by other research groups [Lai96, Pop04, Vasileska10], but to assess and characterize the impact of device self-heating in various technologies, both in the time-zero and time-dependent (i.e. reliability) characteristics, find modulators of this effect and to study the impact of technology-induced geometrical or material modifications.

### *2.6.4 Thermal boundary conditions*

An important aspect of solving differential equations such as the lattice heat equation, are the boundary conditions that are applied. In thermal simulations, typical Dirichlet and Neumann boundaries can be applied. A Dirichlet conditions consist in specifying the value of the solution, i.e. in a thermal

systems this means fixing the in one edge, surface or domain, making that boundary an ideal heat sink.

A Neumann boundary condition consists of imposing the value of the first derivative. This means that this conditions imposes *the flux* of heat through that edge. If a Neumann BC is set to zero, it means that the edge, surface or domain is isolated and no flux of heat enters or outputs that region.

In typical thermal simulations, Dirichlet boundary conditions are applied to either all the outer edges of the simulated domain, or to specific regions, such as the bulk, the contacts and the gate. In Chapter 5, we will show that the location of the boundary conditions can have a distinct effect on the simulation outcome of the temperature profile and average temperature of the device.

### 2.6.5 Self-heating measurement techniques

Electrical assessment of SHE can be mainly achieved in two ways. Either the device's *self-heating induced drive current change* is assessed, which we will call the time-domain techniques; or the *temperature is measured directly*, which we call the direct measurement techniques.

In time-domain techniques, the self-heating effect is "disabled" by measuring the drive current (or its first derivative) at extremely short timescales. Typical examples are pulsed-IV (PIV) [Beppu13], AC-conductance (ACC) [Tenbroeck96] and RF-measurements [Scholten09]. The techniques rely on the fact that at high frequencies, the channel temperature does not follow voltage oscillations, due to the thermal inertia of device, substrate and package. The drive current is reduced due to lower carrier mobility at elevated temperatures and vice versa. Mobility is one of the three main factors (the other ones are threshold voltage $V_{TH}$ and saturation velocity $v_{sat}$) resulting in the overall FET temperature behavior.

The transistor mobility has a very complex temperature dependence, defined by the interplay of the multiple electron scattering parameters: phonon scattering, surface roughness scattering and coulombic scattering. Theoretical calculations indicate that the mobility in non-polar semiconductors, such as silicon and germanium, is dominated by *acoustic* phonon scattering. The resulting mobility around typical operating temperatures is then expected to

be proportional to ~$T^{-3/2}$ whereas it is expected to be proportional to ~$T^{-1/2}$ when dominated by *optical* phonon scattering [Sze81].

It should be noted that each of these time-domain techniques has disadvantages. The major issues with these indirect measurement techniques are that the interpretation of the data is not straightforward and the SHE cannot be completely disabled due to timing constraints of the heating.

### 2.6.5.1 *Pulsed-IV measurements*

The principle of PIV measurements is depicted in Fig. 62. Short voltage pulses are applied to the gate or the drain terminal and the resulting current is measured in 70 to 90% interval of the top side pulse. Presuming that the device is not yet completely heated during the pulse, the drive current will thus overshoot the steady-state current during the pulse.

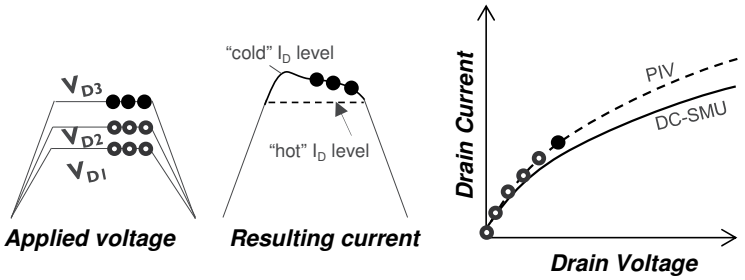



**Fig. 62: Illustration of the concept of pulsed-IV measurements. Short pulses are applied and the current is measured on the top of the pulse. A distinct current overshoot is expected for the period the FET is cold, resulting in a higher effective current w.r.t. DC-measurements.**

An intrinsic limitation of this technique is that only a fraction of the heating is disabled. As the devices continue to scale down, the channel heating time constant:

$$\tau_{channel} = R_{TH}.C_{TH} \tag{2.31}$$

is expected to decrease because the $R_{TH}$ is mostly proportional to the silicon *surface* available for dissipation near the channel, whereas $C_{TH}$ is mainly

determined by the available silicon *volume*. For nanoscale transistors, this implies that the required measurement speed to capture the full time-domain effect can only be obtained by RF techniques [Scholten09].

### 2.6.5.2 *RF-measurements*

In RF-measurements, the output conductance of the device is extracted from S-parameters, measured with a vector network analyzer (VNA), calibrated and using dedicated de-embedding structures [Makovejev11]. The device's output conductance can then be measured over a very wide frequency range. This technique is based on the assumption that, at high frequencies, the channel temperature does not follow voltage oscillations and hence dynamic self-heating is removed. It should be noted that the static part of self-heating, dependent on the DC-operating conditions of the device, are *not* removed.

The conductance difference at low and high frequencies, as illustrated in Fig. 63, (where dynamic self-heating is removed) can then be translated into the $R_{TH}$. Typical frequencies used originally in this technique are in the kHz to MHz range. In advanced nonplanar devices, with a decreased volume-to-surface ratio as discussed above, leading to smaller time constants, higher frequencies are required to obtain self-heating-free characteristics. It is for that reason that these measurement are typically done on *dedicated* RF-structures.

The real part of the thermal impedance (thermal resistance) is proportional to the conductance difference at low and high frequencies according to [Rinaldi01]

$$g_{dd} - g_{ddT} = Re(Z_{TH}) \frac{\partial I_D}{\partial T_A} (V_G g_{gdT} + V_D g_{gddT} + I_D) \tag{2.32}$$

where $g_{dd}$ is the real part of the $Y_{22}$ (which represents the conductance) parameter at low frequency with dynamic self-heating present and $g_{ddT}$ the part at high frequency with dynamic self-heating removed. $Z_{TH}$ is the thermal impedance of the device and $g_{gdT}$ is the real part of the $Y_{12}$ parameter at high frequency.

The difference between the total drain capacitance at low and high frequencies is proportional to the imaginary part of thermal impedance according to

$$C_{dd} - C_{ddT} = \frac{Im(Z_{TH})}{2\pi f}\frac{\partial I_D}{\partial T_A}(V_G g_{gdT} + V_D g_{gddT} + I_D) \tag{2.33}$$

where $C_{dd}$ *and* $C_{ddT}$ are the total drain capacitances extracted from the imaginary part of $Y_{22}$ parameters, respectively, at low frequency with dynamic self-heating present and at high frequency with dynamic self-heating removed, and $f$ is the frequency. The imaginary part of the thermal impedance can then be used to extract $C_{TH}$:

$$C_{TH} = \frac{1}{2\pi f Im(Z_{TH})}. \tag{2.34}$$

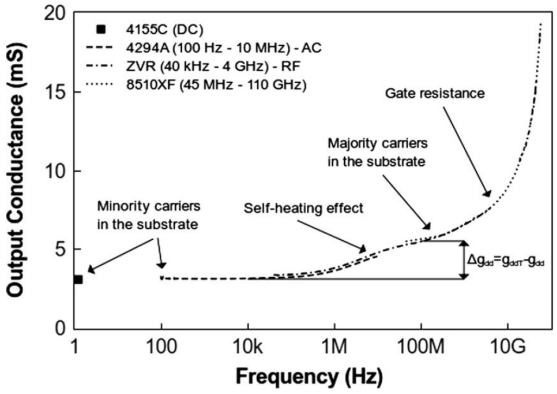



**Fig. 63: Output conductance as function of frequency obtained with different instruments for a SOI FinFET device [replotted from Makovjev11].**

### 2.6.5.3 *Direct measurement techniques*

It should be noted that we cannot use the typical temperature dependent characteristics of the device-under-test (DUT) *itself* to assess its $\Delta T$ during operation. Such characteristics include junction forward bias current or sub-

threshold swing, but in the former case, the current will simultaneously vary due to the potential profile shift in the DUT at varying $V_{DS}$ or $V_{GS}$ conditions rather than due to the SHE, and for the latter, the device will already have cooled down as SS measurements cannot be performed 'on the fly'. Moreover, the device degradation mechanisms as BTI or CHC can be activated when the devices are biased into saturation, and will cause on its turn $V_{TH}$ shifts by oxide charge trapping or sub-threshold degradation by generation of interface states. It is not possible to de-convolute this additional degradation from self-heating.

**Table III: Overview of typically used measurement techniques for self-heating assessment**

| Measurement Technique | Direct or indirect | Observation |
|---|---|---|
| Transistor sub-threshold swing | Direct (ΔT) | Not self-heating regime |
| Junction forward bias current | Direct (ΔT) | Affects device's $I_D$ |
| Gate resistance thermometry | Direct (ΔT) | Special structure required |
| Diode structure | Direct (ΔT) | - Distance critical<br>- Special structure required |
| Pulsed-IV measurement | Indirect (%ID degrad) | Time critical |
| AC-conductance<br>or<br>S-parameter (RF) | Indirect (%ID degrad) | - Time/frequency critical<br>- Partially disabling SHE<br>- Parasitic gate resistance |

Therefore, methodologies to directly sense the device's $\Delta T$, rely mostly on *dedicated designed structures* in which the temperature dependent characteristic of a nearby structure or device is utilized as a sensor. The most common methodology is to probe the temperature of the DUT's gate by resistance measurements, using a 4-terminal connection, called gate resistance thermometry [Su94]. The gate material is selected to feature a strong temperature dependent resistance to increase the measurement resolution. The resistance change in the gate, induced by $\Delta T$ is averaged over the entire gate length, width and depth, which consequently necessitates rigorous calibration.

We will utilize and elaborate on this measurement methodology in Chapter 6. An overview of typically used measurement techniques for self-heating assessment is given in Table III.

## 2.7 Conclusions

In this Chapter, an phenomenological overview and physical insights were given in typical failure mechanisms that are observed in MOSFET devices, with the focus on BTI and self-heating effects.

## 2.8 References

[Alam03] Alam M. A., "A critical examination of the mechanics of dynamic NBTI for pMOSFETs" in Proc. IEDM, pp. 345–348, (2003).

[Ang08] Ang D., Wang S., Du G., and Hu Y., "A Consistent Deep-Level Hole Trapping Model for Negative Bias Temperature Instability," IEEE Trans. Dev. Mat. Rel. , Vol. 8, No. 1, pp. 22–34, (2008).

[Aoulaiche05] Aoulaiche M. et al., "Contribution of fast and slow states to Negative Bias Temperature Instabilities in $HfxSi_{(1-x)}ON$/TaN based pMOSFETs", in Microelectronic Engineering, vol. 80, pp. 134-137, (2005).

[Asenov08] Asenov A. et al., "Advanced simulation of statistical variability and reliability in nano CMOS transistors", in IEEE Proc. IEDM, pp. 1, 2008;

[Bentarzi11] Bentarzi H., "Transport in Metal-Oxide-Semiconductor Structures", Springer, (2011).

[Cho14] Cho M., Bury E., Kaczer B and Groeseneken G., "Hot Carrier degradation in Semiconductor devices" – chapter "Channel Hot Carrier Degradation and Self-Heating Effects in FinFETs", Editor T. Grasser, Springer, (2014).

[Dallman95] Dallmann D.A. and Shenai K., "Scaling Constraints Imposed by Self-Heating in Sub-micron SO1 MOSFET's", IEEE Transactions on Electron Devices, Vol. 42, No. 3, pp. 489-496, (1995).

[Deal67] Deal B. E. et al., "Characterization of the Surface-State Charge (Qss) of the thermally Oxidized Silicon", Journal of Electrochemical Society, Solid State Science, Vol. 114, No.3 , pp. 266-274, (1967).

[Deal80] Deal B.E. et al., "Standardized Terminology for Oxide Charges Associated with Thermally Oxidized Silicon", in IEEE Transactions on Electron Devices, Vol. 27, Iss.3, pp. 606-608(1980)

[Dobkin03] Dobkin D.M., "Principles of Chemical Vapor Deposition", Springer, (2003).

[Ershov03] Ershov M. et al, "Dynamic recovery of negative bias temperature instability in p-type metal-oxide-semiconductor field-effect transistor", Journal of Applied Physics, Vol. 83, No. 8, pp. 1647-1649, (2003).

[Fiegna08] Fiegna C., Yang Y. ,Sangiorgi E., and O'Neill A. G., "Analysis of Self-Heating Effects in Ultrathin-Body SOI MOSFETs by Device Simulation, IEEE Transactions on Electron Devices, Vol. 55, No. 1, pp. 233-244, (2008).

[Fischetti95] M.V. Fischetti and S.E. Laux, "Monte Carlo study of sub-band-gap impact ionization in small Silicon field-effect transistors", in Proc. IEDM, pp. 305-308, (1995).

[Franco12] Franco J., "Reliability of High Mobility (Si)Ge Channel pMOSFETs for Future CMOS Applications", PhD Dissertation, KU Leuven, (2012).

[Franco13] Franco J. et al., "Understanding the suppressed charge trapping in relaxed- and strained-Ge/SiO2/HfO2 pMOSFETs and implications for the screening of alternative high-mobility substrate/dielectric CMOS gate stacks", in Proc. IEDM, pp. 397-400, (2013).

[Garros10] Garros X. et al., "Reliability concerns in High-K/Metal gate technologies", in Proc. ICIDT, pp. 90-93, (2010).

[Giles15] Giles M.D. et al., "High sigma measurement of random threshold voltage variation in 14nm Logic FinFET technology", in Proc. VLSI, pp. 150-151, (2015).

[Grasser07] Grasser T. et al., "Simultaneous extraction of recoverable and permanent components contributing to bias temperature instability", in Proc. IRPS, pp. 801-804, (2007).

[Grasser07a] ] Grasser T. et al., "The universality of NBTI relaxation and its implications for modeling and characterization," in Proc. IRPS, pp. 268–280, (2007).

[Grasser09b] ] Grasser T. et al., "Switching Oxide Traps as the Missing Link between Negative Bias Temperature Instability and Random Telegraph Noise", in Proc. IEDM, pp. 729–732, (2009).

[Grasser10b] Grasser T., et al., "The time dependent defect spectroscopy (TDDS) for the characterization of the bias temperature instability", in Proc. IRPS, pp. 16-25, (2010).

[Grasser10a] Grasser T. et al., "Recent advances in understanding the bias temperature instability", in IEEE Proc. IEDM, pp. 4.4.1-4.4.4, (2010).

[Grasser11] Grasser T. et al, "The Paradigm shift in Understand the Bias Temperature Instability: Fro Reaction-Diffusion to Switching Oxide Traps", in IEEE TED, Vol. 58, No. 11, pp. 3652-3665, (2011).

[Grasser12] Grasser T. et al, "On the Microscopic Origin of the Frequency Dependence of Hole Trapping in pMOSFETs", in Proc. IEDM, pp. 470-473, (2012).

[Guerin09] Guérin C, Huard V. and Bravaix A, "General framework about defect creation at the $Si/SiO_2$ interface", in Journal. Appl. Phys., vol. 105, 144516, (2009).

[Hu85] Hu. C et al., "Hot-Electron Induced MOSFET Degradation-Model, Monitor, Improvement", IEEE Trans. Electron Devices, vol. 32, pp. 375-385, (1985).

[Huard06] Huard V., Denais M. and Parthasarathy C., "NBTI degradation: from physical mechanism to modeling", in Microelectronic Reliability, Vol. 46, No. 1, pp. 1-23, (2006).

[Huard08] Huard V et al., "NBTI degradation: From Transistor to SRAM Arrays," in Proc. IRPS, pp. 289, (2008).

[Huff05] Huff H. and Gilmer D., "High Dielectric Constant Materials, VLSI MOSFET Applications", Springer, (2005).

[Jeppson77] Jeppson K.O. and Svensson C.M., "Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices", in J. Appl. Phys., vol. 48, no. 5, pp. 2004-2014, (1977).

[Kaczer02] Kaczer B., et al., "Observation of hot-carrier-induced nFET gate-oxide breakdown in dynamically stressed CMOS circuits", in Proc. IEDM, pp. 171-174, (2002).

[Kaczer05] Kaczer B. et al., "Disorder-controlled-kinetics Model for Negative Bias Temperature Instability and its Experimental Verification", in Proc. IRPS, pp. 381-387, (2005).

[Kaczer08] Kaczer B. et al., "Ubiquitous Relaxation in BTI Stressing—New Evaluation and Insights", in Proc. IRPS, pp. 20-27, (2008).

[Kaczer10] Kaczer B. et al., "Origin of NBTI Variability in Deeply Scaled pFETs", in Proc. IRPS, pp. 26-32, (2010).

[Kaczer15] Kaczer B., et al., "Extraction of The Random Component of Time-Dependent Variability Using Matched Pairs", published in IEEE Electron Dev. Lett. Vol. 36, Iss. 6, pp. 300-320, (2015).

[Kauerauf13] Kauerauf T., "TDDB in sub-1nm HK/MG gate dielectrics", IEEE IRPS Short Course, (2013).

[Kerber08] Kerber A. et al., "Characterization of Fast Relaxation During BTI Stress in Conventional and Advanced CMOS Devices with HfO2/TiN Gate Stacks", IEEE Transaction on Electron Devices, Vol. 55, No. 11, pp. 3175, (2008).

[Kerber13] Kerber A. et al., "Challenges in the characterization and modeling of BTI induced variability in Metal Gate / High-k CMOS technologies", in Proc. IRPS, pp. 2D.4.1-2D.4.6, (2013).

[Kuhn09] Kuhn K. et al. "Moore's Law Past 32nm: the Challenges in Physics and Technology Scaling" in SSDM, pp. 1-6, (2009).

[Lacaita91] Lacaita A., "Why the effective temperature of the hot electron tail approaches the lattice temperature", in App. Phys. Lett., vol. 59, no. 13, pp. 1623-1625, (1991).

[Lai96] Lai J. and Majumdar A., "Concurrent thermal and electrical modeling of submicrometer silicon devices", J. Appl. Phys. , Vol. 79, pp. 7353 (1996).

[Liu05] Liu Y., "Study of Oxide Breakdown, Hot Carrier and NBTI Effects on MOS Device And Circuit Reliabilty", PhD Dissertation, University of Central Florida, (2005).

[Liu06] Liu W. and Asheghi, M., "Thermal Conductivity of Ultra-Thin Single Crystal Silicon Layers," Journal of Heat Transfer, Vol. 128 pp. 75-83, (2006).

[Lyu97] Lyu J. et al., "Reduction of Hot-Carrier Generation in 0.1um Recessed Channel nMOSFET with Laterally Graded Channel Doping Profile", in IEEE Electron Device Lett., vol. 18, no. 11, pp. 535-537, (1997).

[Mahapatra05] Mahapatra S. et al., "Negative bias temperature instability in CMOS devices", in Microelectronic Engineering, vol. 80, pp. 114–121, (2005).

[Mahapatra09] Mahapatra S. et al., "S. Mahapatra, V. D. Maheta, A. E. Islam, and M. A. Alam, "Isolation of NBTI stress generated interface trap and hole-trapping components in PNO p-MOSFETs," in IEEE Trans. Electron Devices, vol. 56, no. 2, pp. 236–242, (2009).

[Makovejev11] Makovejev S.,Olsen S., and Raskin J.-P., "RF Extraction of Self-Heating Effects in FinFETs", IEEE Transactions on Electron Devices, Vol. 58, No. 10, pp. 3335-3341, (2011).

[Mitani05] Mitani, Y., Satake, H. and Toriumi, A. "Negative Bias Temperature Instability in Ultra-Thin SiON", in Electrochemical Society, Vol. 2005, No. .1, pp. 340-355, (2005).

[Nicollian71] Nicollian E. et al., "Electrochemical charging of thermal SiO2 films injected electron currents", Journal of Applied Physics, Vol. 42, No. 13, pp. 5654-5664, (1971).

[Ogura80] Ogura S. et al., "Design and Characteristics of the Lightly Doped Drain-Source (LDD) Insulated Gate Field-Effect Transistor", in IEEE Trans. Electron Devices, vol. 27, no. 8, pp. 1359-1367, (1980).

[Ong90] Ong T.-C., Ko P.K. and Hu C. "Modeling of substrate current in p-MOSFET's", in IEEE Electron Dev. Lett., vol. 8, pp. 413-416, (1987).

[Pantelides07] Pantelides S.T. et al.," Hydrogen in MOSFETs – A primary agent of reliability issues", in Microelectronics Reliability, vol. 47, pp. 903-911, (2007)

[Pelgrom89] Pelgrom M., Duinmaijer A., and Welbers A., "Matching Properties of MOS Transistors", in IEEE JSSC, vol. 24, pp. 1433-1440, (1989).

[Pop04] Pop. E, "Self-Heating and scaling of thin-body transistors", PhD Dissertation, Stanford University, (2004).

[Rauch07] Rauch S.E. et al., "Review and Reexamination of Reliability Effects Related to NBTI Statistical Variations", IEEE Trans on Dev. Mat. Rel. 7, pp. 524, (2007).

[Rhyner13] Rhyner R., and Luisier M., "Self-heating effects in ultra-scaled Si nanowire transistors", IEEE International Electron Devices Meeting, pp. 790-793, (2013).

[Schroder05] Schroder D.K. "Negative bias temperature instability: Physics, materials, process, and circuit issues" available at http://www.ewh.ieee.org/r5/denver/sscs/Presentations/2005.08.Schroder.pdf, (2005).

[Schroder07] Schroder D.K. "Negative bias temperature instability: What do we understand?", in Microelectronics Reliability, vol. 47, pp. 841-852, (2007).

[Shockley61] Shockley W. et al., "Problems related to p- n junctions in silicon", in Solid State Electronics 2, pp. 35-67, (1961).

[Snow65] Snow E. H. et al., "Ion Transport Phenomena in Insulating Films", Journal of Applied Physics, Vol. 36, No. 5, pp. 1664-1673, (1965).

[Takeda83] Takeda E, Suzuki N. and Hagiwara T., "Device Performance Degradation to Hot-Carrier Injection at Energies Below the Si-$SiO_2$ Energy Barrier", in Proc. IEDM, pp. 396-399, 1983.

[Takeuchi07] Takeuchi M. et al., "Understanding Random Threshold Voltage Fluctuation by Comparing Multiple Fabs and Technologies" in Proc. IEDM., pp. 467-470, (2007).

[Teo09] Teo Z.Q., Ang D.S. and See K.S., "Can the Reaction-Diffusion Model Explain Generation and Recovery of Interface States Contributing to NBTI?", in Proc. IEDM, pp. 737-740, (2009).

[Toledano11] Toledano-Luque M. et al., "From mean values to distributions of BTI lifetime of deeply scaled FETs through atomistic understanding of the degradation", in Proc. VLSI, pp. 152-153, (2011).

[Townsend10] Townsend J. S.E, "The theory of ionization of gases by collision", Constable, Londen, (1910).

[Triebl12] Triebl O., "Reliability Issues in High-Voltage Semiconductor Devices", PhD Dissertation, TU Wien, (2012).

[Vasileska10] Vasileska D. ,Raleva K. and Goodnick S. M., "Electrothermal Studies of FD SOI Devices That Utilize a New Theoretical Model for the Temperature and Thickness Dependence of the Thermal Conductivity, IEEE Transactions on Electron Devices", Vol. 57, pp. 726-728 (2010).

[Wallace05] Wallace R. M. and Wilk G. D., "Materials Issues for High-k Gate Dielectric Selection and Integration", in High Dielectric Constant Materials, Springer Series in Advanced Microelectronics Vol. 16, pp. 253-286 (2005).

[Weber88] Weber W. et al., "Dynamic stress experiments for understanding hot-carrier degradation phenomena,"in IEEE Trans. Electron Devices, vol. 35, pp. 1476–1486, (1988).

[Wiedemann53] Wiedemann G, Franz R.: "Ueber die Wärme-Leitungsfähigkeit der Metalle", in Ann. Phys. 89, No. 8, pp. 497–531 (1853).

[Ziman60] Ziman J.M., "Electrons and Phonons", Clarendon Press, Oxford, (1960)

# Chapter 3: Characterizing BTI-reliability in ultra-thin EOT devices

Based upon a thorough understanding of the device Bias Temperature Instabilities (BTI) by extensive measurements which were enabled by a newly introduced measurement technique, conclusions are drawn on their dependencies on process variations in the gate stack and we provide guidelines to process engineers to mitigate BTI degradation in UT-EOT devices.

## 3.1 Introduction

Gate dielectric reliability is a well-documented area of research, with literature spanning over multiple decades discussed the $SiO_2$ reliability, and in the last decade also high-k reliability. The introduction of HKMG devices had required a substantial improvement in the understanding of the transistor instability, particularly in PBTI and NBTI.

Previous experimental works reporting on UT-EOT stacks indicated that reliability margins tended to decrease rapidly and become a major obstacle for EOT scaling in devices. [Kerber09, Groeseneken10]. Thus, understanding and modeling of the device instabilities and their dependencies on process variations becomes a crucial issue for semiconductor industry. The exact mechanisms that lead to the *enhanced* NBTI, specifically for HKMG are however not well understood.

Inherent to the techniques to obtain UT-EOT, is that slight changes in processing conditions, can strongly impact device parameters, such as EOT, $V_{TH}$, leakage current, and also reliability. Therefore, it is unknown if this decrease in BTI resilience is either *fundamental* or *processing related*. To investigate this, systematic benchmarking of UT-EOT devices and various process "recipes" is essential.

The reliability of these UT-EOT devices could, up to now, only be extracted on transistors which are only produced if the processing conditions are established and therefore BTI data on UT-EOT stacks is scarce.

In this Chapter, we will first introduce the various process methodologies to obtain industry-relevant UT-EOT devices. Then, we will study and describe which defects are responsible for charge trapping during BTI stress conditions. Then, we will show how plain capacitors can be used for BTI evaluation, comparable to typical transistor measurements by accessing the same defects. We will discuss the impact of correct electric field extraction and prove the validity of our BTI-evaluation technique on various UT-EOT gate stacks and how measurement artefacts can be resolved. We will then apply this methodology to assess the BTI reliability for various processes. Based upon the measurement results of capacitor lot, two hypotheses are proposed that can explain the accelerated BTI degradation. Finally, we will propose alternatives for C-V extraction in leaky and nanoscale devices.

## 3.2 Processing options for UT-EOT devices

In scaling the gate oxide thickness, one of the benchmarks to qualify the high-k gate stack is the EOT. The EOT is defined as follows:

$$EOT = t_{SiO_2} + t_{high-k}\frac{k_{SiO_2}}{k_{high-k}} \tag{3.1}$$

with $t_{SiO2}$ and $t_{high\text{-}k}$ the thickness of the interfacial $SiO_2$ and high-k layers respectively, and the k values are their respective relative dielectric constants. For $SiO_2$ this constant is 3.9, whereas for high-k this dielectric thickness depends on the crystal structure or the amorphysation degree of the layer. The EOT is measurable by C-V measurements as it is inversely proportional to the saturation capacitance in accumulation.

### *3.2.1 Scavenging to obtain ultra-thin EOT devices*

Over the last years, the improvements in CMOS device performance via gate oxide scaling have been achieved in gate-first (GF) integration schemes through a metal-inserted poly-Si stack (MIPS) incorporating process flow and by using interfacial layer scavenging [Ragnarsson09, Huang09]. *Oxygen scavenging* is the best known technique to produce a zero interfacial layer

(zero-IL) $HfO_2$ device. It was shown that the selection of the metal electrode has a significant effect on the electrical and chemical properties of the gate stack. For example, it was found that the use of Ti as a metal gate prevents the growth of an interface layer in the Si/high-k interface, and can even reduce an existing interface layer [Kim04,Choi10]. It is this phenomenon which is known as the "scavenging" effect. The common mechanism used to explain this phenomenon was suggested by Ando *et al.* by the diffusion of oxygen into (direct scavenging) the high-k dielectric or through (remote scavenging) the high-k dielectric into the metal electrode [Ando09]. Direct-scavenging schemes thus incorporate the scavenging elements within the high-κ layers, whereas the remote-scavenging schemes isolate the scavenging elements from the high-κ layers.

The reaction can be expressed as:

$$Si + \frac{2}{y} M_x O_y \leftrightarrow \frac{2x}{y} M + SiO_2 \tag{3.2}$$

where M is typically the dopant element in the metal gate, such as La as depicted in Fig. 64 below.

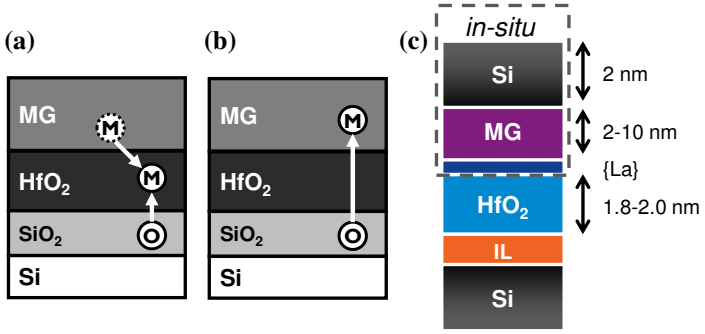
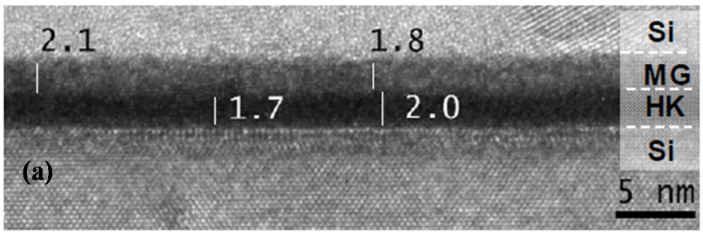



**Fig. 64: Illustration of (a) direct and (b) remote interfacial layer (IL) scavenging. In direct scavenging the metal ion can be *already* present in the high-k or be in-diffused. In the case of remote scavenging, the scavenging metal is isolated from the high-k material and the reaction takes place in the metal gate. (c) Example of a gate stack with in-situ Si cap deposition to prevent regrowth of the IL during the scavenging activation anneal.**

The typical process flow for gate stack deposition studied in this work, is as follows:

- in the cleaning step, a ~1 nm thick very pure oxide IL is grown directly on Si.
- afterwards, 1.8-2 nm $HfO_2$ layer is deposited via atomic layer deposition (ALD),
- next an in-situ physical vapor deposited (PVD) electrode is grown, containing a La/Al cap and a 2 up to 10 nm TiN or TaN metal gate,
- and finally, a Si-cap is in-situ formed on top.




| (b) Assumption | $t_{high-\kappa}$(Å) | $t_{IL}$(Å) | κ-value |
|---|---|---|---|
| Zero interface | 18 | 0 ⇨ | 15 |
| high-κ permittivity=20 | 18 | 1 ⇦ | 20 |

**Fig. 65: (a) HRTEM of a Metal-Inserted-poly-Si (MIPS) device with a 2nm $HfO_2$, 2nm Metal Gate (MG) and 2nm Si. No interfacial SiO2 layer is observed. Indicated thicknesses are in nm [replotted from Ragnarsson09] and (b) Illustration of how an assumption on the interfacial layer thickness is needed to determine the dielectric constant or vice versa.**

The key factor for thin EOT is preventing additional oxygen (in form of $H_2O$/OH) to reach the $HfO_2$ or IL, since otherwise IL regrowth happens during anneal [Ragnarsson09]. A high-resolution TEM (HRTEM) of such a 'zero-IL' gate stack is shown in Fig. 65. Note however that even though no interfacial layer is observed on the HRTEM, it is impossible to claim with certainty the physical thicknesses of high-k and $SiO_2$ IL (if existing), due to the large uncertainty of the k-value of those thin-film $HfO_2$ films. The physical

thickness of the high-k layer typically is a measureable unit, and can be estimated from the number of atomic layer deposited (ALD) $HfO_2$ cycles.

### 3.2.2 Gate-first versus gate-last integration

At the time of the study, a trade-off existed between the GF and the replacement gate (RMG), typically with high-k last (HKL). Whereas the GF technique has the advantages of a more simple processing, the gate stack has to undergo the thermal budget for subsequent steps used for junction activation et cetera. During these steps, the dielectric might suffer from additional defect generation and oxide charging due to reaction with the Poly Si. This causes the $V_{TH}$ to roll-off. The RMG does not suffer from these issues as the final gate stack is deposited after junction formation. The subsequent thermal treatment is typically more moderated, believed to have positive impact on the $V_{TH}$ roll-off. However, the process integration complexity increases and the RMG process has inherent scaling issues. As the subsequent filler material is inserted in the opened gate-trench, both the bottom and the sidewalls are covered. For short gate lengths, little volume is remaining in the gate trench after high-k and WF metal filling. As such, the gate resistance might increase. Integration schemes of both techniques are depicted in Fig. 66. A GF process using a Metal-Inserted-Poly-Si (MIPS) technique is shown. In this technique, first the high-k and the gate metal (GM) for nFET (or pFET) is deposited. After patterning the pFET (or nFET) regions, the corresponding gate stack is deposited. After another patterning/etching step, only the gates remain. In a final stage, the source/drains and the contacts are formed and the interlayer dielectric (ILD) is deposited.

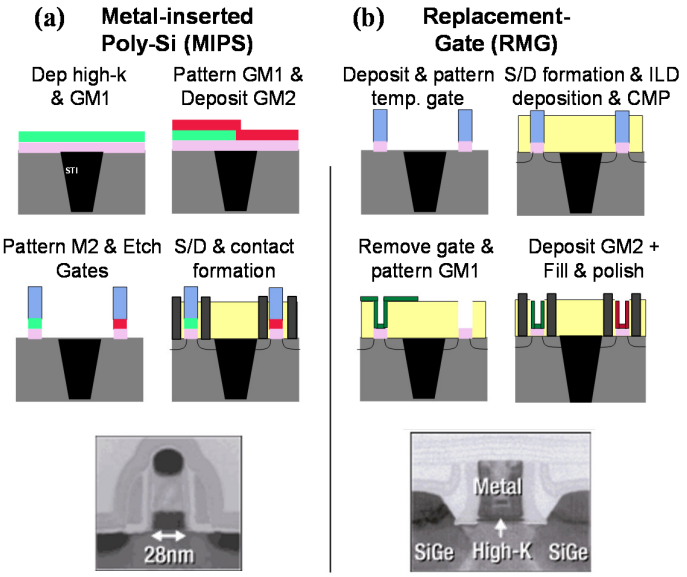



**Fig. 66: Illustration of (a) Metal-Inserted-Poly-Si (MIPS) process, typically referred to as Gate-First (GF) and (b) replacement gate (RMG) process, referred to as Gate-Last (GL).**

Fig. 66 also shows the replacement gate (RMG) process, referred to as Gate-Last (GL). In the GL option, *dummy gates are patterned first*. After S/D formation and ILD deposition and chemical-mechanical polishing (CMP) step, the dummy gates are removed and the corresponding gate metals are deposited. Note that the high-k dielectric can be already be deposited *before* ("high-k first") or *after* the dummy gate removal ("high-k last").

Fully-silicided Gate (FUSI) process is not discussed here, as it was already phased-out at the time of the study.

## 3.3 Access to defect bands during BTI evaluation

Despite earlier contradictory results [Ho2012, Huard03], there is now a consensus that not only the creation of interface traps, but also the trapping of positive charges (holes) in the bulk oxide plays an important and even dominant role during (N)BTI stress.

However, with the scaling towards ultra-thin layers, the boundary between interface and bulk defects becomes vaguer. This is caused by the disappearance of the interfacial $SiO_2$ layer, and the corresponding change in the chemical interface quality. A suggestion that we will verify later in this Chapter, that the poor properties are related to high-k defects (and/or their accessibility) rather than defects to in the interfacial $SiO_x$-like layers.

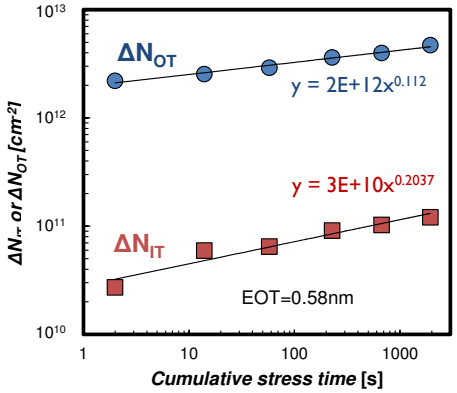



**Fig. 67: Increase in bulk and interfacial charge trapping due to BTI stress in ultra-thin EOT (UT-EOT) devices. The interfacial charge is measured by ring-oscillator charge pumping (ROCP) at 500MHz. The bulk charge trapping is clearly dominant over the interfacial charge trapping [replotted from Cho11].**

The increase in oxide trapping and interface trapping (quantified by ring-oscillator charge pumping) after BTI stress in UT-EOT devices was shown by Cho et al. (Fig. 67). [Cho2011]. The bulk trapping is about 2 orders of magnitude higher than the interface trap generation. This is thus an indication

that we need to focus on the bulk oxide trapping if we want to optimize the BTI reliability of these UT-EOT stacks.

### 3.3.1 Charge trapping in the oxide during BTI stress

In this Section, we will show how the extended Measure-Stress-Measure (eMSM) is scanning the defect band(s) in $Si/SiO_2/HfO_2$ systems.

During stress, holes from the inversion channel in a pFET (n-substrate) are injected in the bulk high-k and in the interfacial $SiO_2$. In equilibrium, the hole traps above the fermi level are filled, whereas the hole traps below the fermi level are empty of holes. Based on the textbook equations [Sze81], the offset of the injection point of the channel carriers with respect to the top of the dielectric conduction band $\zeta_{defect\_inv}$, can be described as follows:

$$\zeta_{defect\_inv} = \chi_{Si} - \chi_{SiO_2} + \frac{E_g(Si)}{2} - \phi_f + \Psi_s \tag{3.3}$$

with $\chi_{Si}$ and $\chi_{SiO2}$ the conduction band-offsets of Si and SiO2 respectively, $E_g(Si)$ the silicon bandgap, $\phi_f$ the Fermi-level and $\Psi_S$ the band bending in the silicon. This offset energy can be rewritten as:

$$\zeta_{defect\_inv} = \phi_{barrier} + \frac{E_g(Si)}{2} + kT\ln\left(\frac{N_D}{n_i}\right) \tag{3.4}$$

with $\phi_{barrier}$ the $Si/SiO_2$ barrier energy, $N_D$ the donor dopant concentration and $n_i$ the intrinsic Si carrier concentration and is illustrated in Fig. 68.

During the relaxation, the measurement voltage $V_{MEAS} \sim V_{TH}$ minimizes influences of both subthreshold slope and mobility changes [Grasser07]. Other factors potentially influencing the FET current during measurement include changes in gate oxide leakage (SILC and soft breakdown) and in drain junction leakage (since $V_D \neq 0V$). These factors are monitored post-stress by a full $I_D$-$V_G$ sweep to ensure there was no influence during the test.

As can be seen from the band diagram in Fig. 68, when increasing the stress voltage, the electric field in the oxide will increase and more holes will

become accessible for charge trapping. In recent work, it was shown by Franco *et al*. that there are strong indications that this defect distribution for holes is singular and in first-order Normal-distributed for typical $Si/SiO_2/HfO_2$ systems [Franco14].

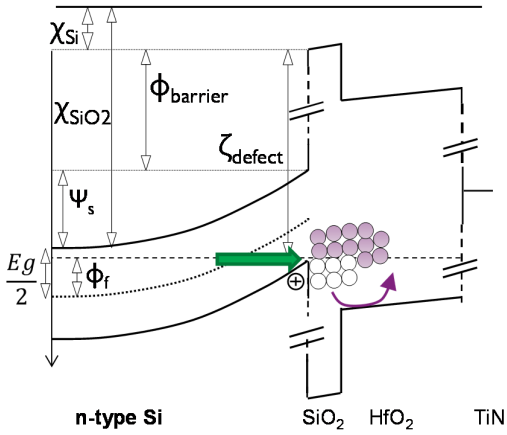



**Fig. 68: Holes from the inversion channel in a n-substrate Si pFET are injected in the bulk high-k and in the interfacial $SiO_2$ layer during NBTI stress. In equilibrium, the hole traps above the Fermi level are filled, whereas the hole traps below the fermi level are empty.**

## 3.4 Capacitors for BTI reliability assessment

In this Section, we will elaborate on the use of plain capacitors for short-loop reliability assessment. Capacitors lots in a fab-environment have typical turn-around times of only a few weeks, whereas a transistor lot takes months to fabricate. As the cost of a lot is directly related to the number of 'lot-turns', i.e. the number of steps in the semiconductor fab, the cost for the production of a capacitor lot is also strongly reduced over a transistor lot.

As we want to study intrinsic and extrinsic gate stack properties, mainly for bulk defect reduction or defect band shifting, many varying recipes are required to fulfill this study.

For this reason, we will elaborate a technique to quantitatively assess the BTI reliability of capacitor lots, and verify is this technique yields results that are representative with respect to fully-processed transistor devices. To do so, we first look if how we are able to quantify charge trapping with *C-V* measurements and how these results relate to typical $I_D$-$V_G$ measurements.

### *3.4.1 Quantifying charge trapping with C-V*

Tracking of $\Delta V_{TH}$ by capacitance measurements and secondly, the equivalence of $V_{TH}$ and flat band voltage ($V_{FB}$) shifts has been shown before in memory devices [Toledano11] and in FET devices [Ando11]. In Fig. 69, we show how trapped charged in a large-area FET can be measured as a characteristic $V_{TH}$ and $V_{FB}$ shift in the *C-V* curve.

In plain capacitors however, i.e. devices with no source/drain junctions and consequently no inversion layer in high-frequency *C-V* (HF-CV), due to the lack of (minority) carrier generation, oxide charge trapping can *only* be observed as a shift of the accumulation region, quantified by $\Delta V_{FB}$.

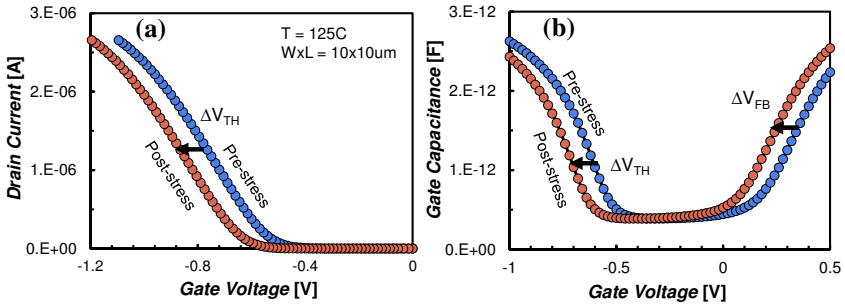



**Fig. 69: (a) The trapped charges in a FET can be measured as a characteristic $V_{TH}$ shift in a I-V measurement post-stress or (b) as $V_{TH}$ and $V_{FB}$ shifts of the *C-V*.**

An important assumption that we made, is that the observed $\Delta V_{TH}$ corresponds to the observed $\Delta V_{FB}$. In the next Section, we will explain why this is (not) the case (for nanoscale devices).

### 3.4.2 Impact of current percolation on $\Delta V_{TH}$ and $\Delta V_{FB}$

Considering the scenario of a 10x10um² transistor with an inversion thickness $t_{inv}$ of 1 nm. The average impact per trap $\eta_0$ based on the simple charge-sheet approximation then is $4.6\text{x}10^{-5}$ mV. To obtain a 50 mV $\Delta V_{TH}$ or $\Delta V_{FB}$, in the order of $1\text{x}10^{6}$ charges are thus involved.

A virtue of the high amount of oxide charges involved in these large transistors, is that the current flows quasi-continuously (i.e. not percolated) from source to drain. This is an important consideration to make, since in nanoscale devices (i.e. in the case of percolated current), the *mean* $\Delta V_{TH}$ *impact per charge* will be higher than the *predicted* mean impact per charge $\Delta V_{TH}$ due to electrostatic screening, *because of this percolated current* [Franco12].

For large device, we can thus safely assume that charges trapped in the dielectric uniformly screen the gate charge uniformly over the *entire device area* and have a ubiquitous gate screening effect over a *wide voltage bias*, in contrast to small devices [Franco12]. Note that according to the same source, the error using the charge sheet approximation for the above calculation is, even in nanoscale devices, in no case larger than 1 order of magnitude.

### 3.4.3 Accessibility of the defect band

In this Section, the accessibility of the same defects or defect levels for both n-type and p-type devices is corroborated.

In order to replicate the standard *I-V*-eMSM sequence for (N)BTI, where (p)MOSFETs are typically stressed in the inversion regime, one could suggest that in capacitors, lacking the inversion regime in HF-CV, the DUTs should therefore be stressed in inversion but subsequently sensed in (positive $V_G$) accumulation regime. This is needed because the *C-V* is flat around $V_{TH}$, which implicates no sensitivity of the measurement tool.

Grasser et al. have however shown that changing the gate polarity throughout stress and relaxation strongly perturbs the trapped charge balance. As a result, more charges are emptied from the defect band during the relaxation, i.e. resulting in a underestimation of $\Delta V_{FB}$ [Grasser11]. Thus, for capacitors, only stressing and measuring both *in accumulation* is the

appropriate manner to acquire charge-induced voltage shifts. It is however a necessary prerequisite that we are accessing (a) the same defects in the gate oxide and (b) stress and sense the oxide under identical fields, to obtain commensurable shifts.

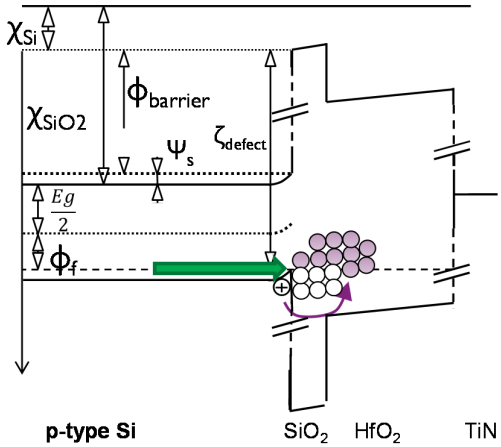



**Fig. 70: Holes from the accumulation layer in a p-substrate Si capacitor are injected in the bulk high-k and in the interfacial $SiO_2$ layer during NBTI stress.**

Also for stressing the devices in accumulation, the band diagram can be drawn (Fig. 70). The band diagram shows how holes from the accumulation layer in a p-substrate Si capacitor are injected in the bulk high-k and in the interfacial SiO2 layer during NBTI stress. In equilibrium, the hole traps above the fermi level are filled, whereas the hole traps below the fermi level are empty. Similar to Eq. 3.4, the injection level for the accumulation carriers with respect to the oxide conduction band can be calculated as:

$$\zeta_{defect\_acc} = \phi_{barrier} + \frac{E_{g\_Si}}{2} + kT \ln\left(\frac{N_A}{n_i}\right) + \psi_s(V_g) \qquad (3.5)$$

with $N_A$ being the number of acceptor dopants in the p-type substrate. Combining Eq. (3.4) and Eq. (3.5), we get:

$$\zeta_{defect\_inv} = \zeta_{defect\_acc} \Leftrightarrow N_A = N_D \quad . \tag{3.6}$$

We thus observe that the injection point of the defects can only be identical in accumulation and in inversion threshold if the net substrate doping level is the same for nFETs and pFETs (but opposite in polarity)

## 3.5 Parameter extraction from C-V measurements

From a correctly obtained *C-V* curve, several MOS parameters can be obtained: dielectric thickness (EOT), threshold ($V_{TH}$) and flat band-voltage ($V_{FB}$), channel doping concentration ($N_D$), inversion ($E_{OX}$) and depletion fields, and gate-doping concentration in case of a doped gate. The *C-V* is also needed for extraction of the device's mobility. Each of the above described parameters will influence the shape of the curve. With decent fitting procedures, numerically solving Poisson and Schrödinger equations (e.g. Hauser's CVC fitting tool [Hauser98]) these parameters can be obtained based on the classical semiconductor equations [Nicollian82], extended with some quantum mechanical corrections, such as taking into account the quantization of the density of states and the shifting of the carrier centroid away from the $Si/SiO_2$ interface.

For ultra-thin EOT devices, the charge centroid is positioned about ~1 nm away from interface, which is seen as ~0.4 nm in the equivalent oxide thickness, due to the difference in dielectric constant between Si and $SiO_2$. This will yield the difference between the extracted *inversion thickness* $t_{inv}$, which quantifies the distance towards the inversion layer and the EOT, which quantifies the equivalent thickness of the dielectric.

As a result, contrary to the thick-oxide situation, the accumulation capacitance does not saturate towards a certain constant value [Ricco88] but keeps increasing with gate accumulation bias. A similar effect will be observed in inversion, where the quantization of the energy levels in the channel can become significant.

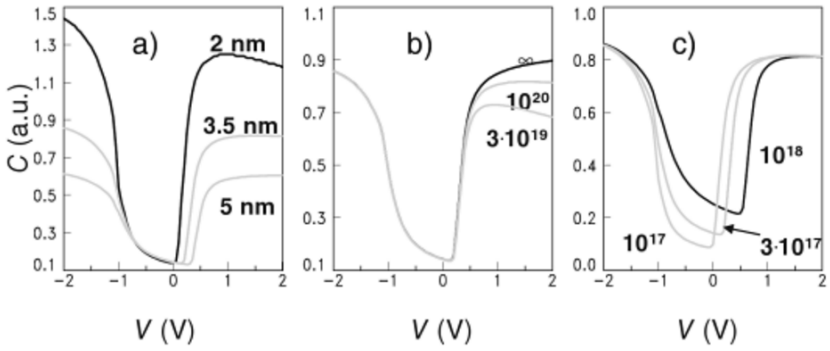



**Fig. 71: Ideal C-V curves as described according to classical MOSFET equations, for (a) various oxide thickness, (b) various gate poly Si doping density and (c) substrate doping density [replotted from Baklanov07].**

Based on (split-)*C-V* measurements (on transistors), we extract essential device parameters such as the EOT, substrate doping, $V_{FB}$ (and $V_{TH}$), shown in Fig. 72. Both the $C_{GA}$-$V_G$ and the $C_{GB}$-$V_G$ were fitted using Hauser's CVC fitting tool [Hauser98]. The fit is shown to excellently describe the experimental data, including the earlier described effects. From this model, the *effective electric field* in the oxide $E_{ox}$ as a function of $V_G$ or $V_{OV}$ can also be calculated.

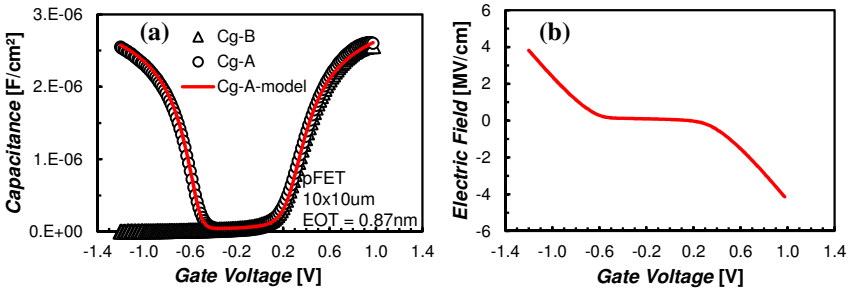



**Fig. 72: (a) Split *C-V*-measurement of a UT-EOT transistor, fitted both the $C_{GA}$-$V_G$ and $C_{GB}$-$V_G$ with Hauser's CVC fitting, yielding essential device parameters such as EOT. (b) The effective electric field $E_{OX}$ in the dielectric as predicted by the model.**

Following the derivation in Eq. (3.5), we change the device polarity. A nFET on the same device lot and identical but opposite substrate doping is stressed at negative bias conditions and $\Delta V_{FB}$ is tracked by *C-V*-eMSM. Fig. 73 compares various techniques to estimate the electric field in the oxide and shows that most commonly used approximations for the electric field are not exact. Typical approximations for the oxide electric field can lead to mistakes up to 30% in lower stress conditions. The only correct approximation is:

$$E_{ox} = \frac{V_G - \Psi_{Si}}{t_{inv}} \tag{3.7}$$

but requires knowledge of the $\Psi_{Si}$, which can by itself, only be determined by capacitance measurement. Therefore, a (split-)*C-V* measurement is thus always a prerequisite in order to obtain the correct oxide electric field.

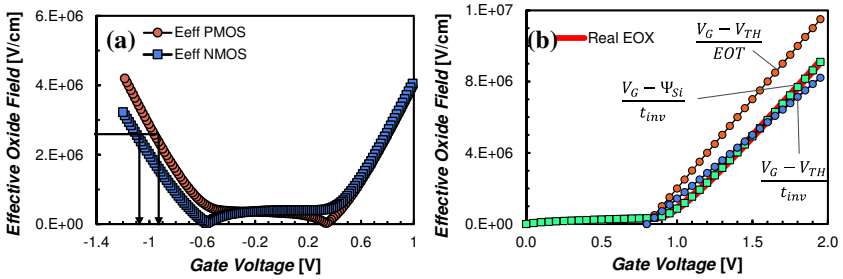



**Fig. 73: (a) Extracted electric field by *C-V*-measurement for a comparable 10x10 nFET and pFETs to determine the equivalent stress voltages. (b) Comparison of various techniques to estimate the electric field in the oxide. Typical approximations for the oxide electric field can lead to mistakes up to 30% in lower stress conditions.**

## 3.6 Development of the CV-eMSM technique

To verify the above described derivation experimentally, we propose the following methodology:

First we will verify and examine the limitations of the LCR setup by firstly measuring typical *I-V*-eMSM $\Delta V_{TH}$ and subsequently tracking $\Delta V_{TH}$ by *C-V*-eMSM for an identical 10x10µm p-channel *transistors*.

After we have confirmed the measurement setup, we will apply the *C-V* MSM on a ***nFET*** of the same dimensions, doping concentration (but opposite in polarity as the pFET doping), and at the same overdrive conditions. If the above derivation holds, we are accessing the same defect band, thus similar degradation should be observed.

In a final stage, we will convert to plain p-substrate capacitor, which has no junctions. Also here, the same reasoning as above holds.

### *3.6.1 Limitations of LCR-based measurements*

This first part of the experiment will examine the impact of the increased latency and the reduced accuracy, both intrinsic weak spots in the LCR setup compared to the SMUs.

In this experimental setup, there is, by definition, no difference in device polarity or stress voltages, except for the difference in applied horizontal field due to the $V_{DS}$ (50 mV for *I-V* versus 0 V for *C-V*). The latter is expected to have negligible impact, since the lateral field is 5 orders of magnitude lower than the transversal field, i.e. 50V/cm compared to 5 MV/cm. Hot carrier degradation experiments have already shown that at low $V_{DS}$, no additional degradation is observed [Spessot14].

Fig. 74 shows that utilizing both methods to assess the $V_{TH}$ degradation result in very similar $V_{TH}$ relaxation trends. It can be observed that the whereas the IV-MSM has below 2ms of delay in the relaxation, this delay increases up to 20ms for the CV-MSM.

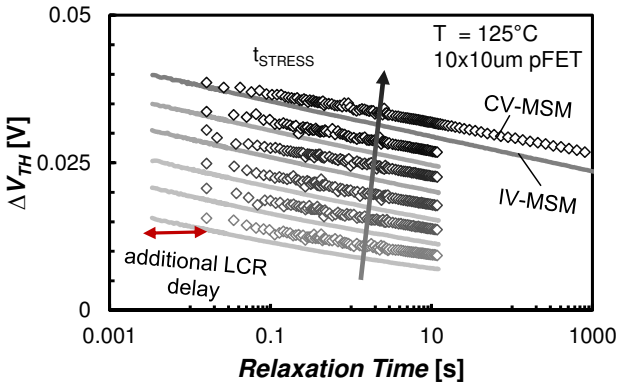



**Fig. 74: The relaxation of the $\Delta V_{TH}$ is tracked for pFET transistor, measured by IV-MSM (solid line) and CV-MSM (symbols). Note the intrinsic higher delay (20ms) of the LCR meter, resulting in fewer decades of measurement results.**

### *3.6.2 CV-MSM measurements on n-type devices*

In Fig. 75, the relaxation of the $\Delta V_{FB}$ measured for a nFET measured with C-V-MSM is compared with the $\Delta V_{TH}$ relaxation observed for a pFET measured with IV-MSM *at identical overdrive conditions* ($V_{OV} = V_{GSTRESS} - V_{TH}$ and $V_{OV} = V_{GSTRESS} - V_{FB}$ for pFET and nFET respectively). The nFET shows slightly lower degradation than the corresponding pFET. If we extract the electric field in the nFET at this overdrive condition based on the CV-extraction described above, we observe that it is only 3.4MV/cm whereas it is 4.2MV/cm for the pFET.

As shown by Cartier et al. and later by Prasad [Cartier11, Prasad13], NBTI is oxide electric field rather than voltage dependent. The reduced $V_{FB}$ shifts are thus due to this lower resulting oxide field, thereby accessing fewer defects. Therefore, in order to guarantee the equivalence between n-type and p-type devices, we have to extract the corresponding overdrive voltage conditions for the pFET and translate it towards an *equivalent oxide electric field* condition during stress.

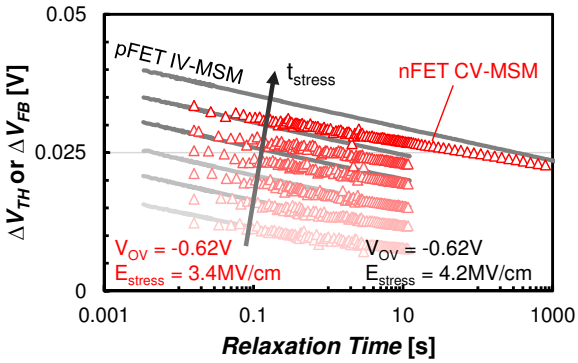



**Fig. 75: (a) The relaxation of the $\Delta V_{FB}$ measured for a nFET measured with CV-MSM is not fully corresponding to the relaxation of the pFET at identical *overdrive* conditions, because of different electric fields.**

In Fig. 76 the relaxation data for the *corresponding capacitor* of the nFET device are shown, i.e. a p-substrate capacitor. In this case, we corrected the $V_{OV}$ to correspond with the pFET stress conditions of 4.2MV/cm.

As a result, it can be seen that the degradation and relaxation of the $V_{TH}$ and $V_{FB}$ for the pFET transistor and p-substrate capacitor are matching.

The experiment was subsequently repeated for various stress conditions and the corresponding BTI lifetime extrapolation trends are extracted. The procedure to do so is as follows: the $\Delta V_{TH}$ and $\Delta V_{FB}$ relaxation transients are fitted with the universal model. From this fit, the degradation at 1 ms of relaxation is extracted.

Then, the typical lifetime extraction procedure is adapted as explained in Chapter 2: for each stress condition, the stress time to reach the degradation criterion is extracted. In this case, a 30 mV $V_{TH}$ shift criterion was utilized. The stress measurement window was 2s up to 1200s.

Finally, a power-law extrapolation of these lifetimes with respect to the stress conditions is performed. As a result of the discussion above, the generated lifetime plot is now plotted as a function of *oxide electric field* $E_{OX}$ instead of $V_{OV}$.

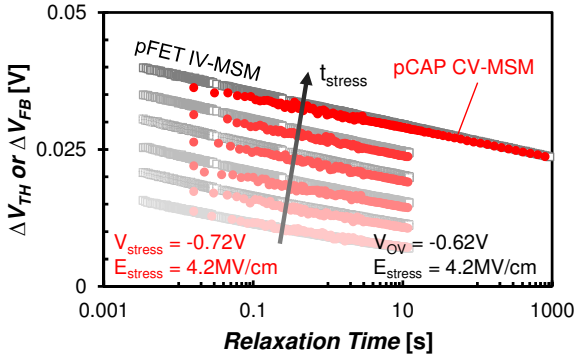



**Fig. 76: The relaxation of the $\Delta V_{FB}$ measured for a p-substrate capacitor measured with *C-V*-MSM corresponds to the relaxation of the pFET (thus n-substrate) transistor at identical electric fields.**

Fig. 77 shows the result of this plot. For all the *I-V* and *C-V*-eMSM cases, this yields an expected *overdrive electric field* of around 1.2 MV/cm, in this case equivalent to an overdrive voltage $V_{OV} = V_G - V_{TH}$ of 0.22 V. This experiment thus confirms *in every step* our proposed verification and thereby confirms the equivalence between *I-V* and *C-V*-eMSM.

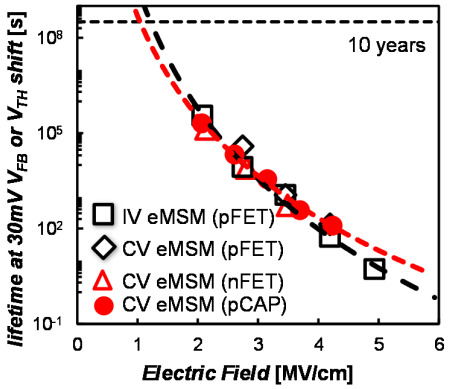



**Fig. 77: BTI lifetime extrapolations of earlier described degradation experiments confirm the consistency as predicted by the theory. Note the low overdrive electric field 1.2MV that is predicted for a 10-years DC operation lifetime.**

The equivalence can be further corroborated in other gate stacks where varying annealing conditions result in different EOT thickness due to modification of the scavenging balance (Eq. 3.2). Lifetime extrapolations comparing the result of the IV-MSM on pFETs and CV-MSM on p-substrate capacitors are shown in Fig. 78 (a). The experimental data shows good agreement between both techniques for each voltage condition. Also the predicted operating lifetimes for each of the gate stacks corresponds.

An important consideration that we should make here, is that the WF metal used for n-type and p-type devices can have different scavenging properties. As a result, the EOT of n-type and p-type devices on the same wafer can slightly vary, as shown in Fig. 78 (b). In this case, this EOT discrepancy between n- and p-type devices is not impacting lifetime prediction. For devices with EOT even <8 Å, were the BTI is strongly dependent on the remaining $SiO_2$ interfacial layer thickness, we expect that this effect might become non-negligible.

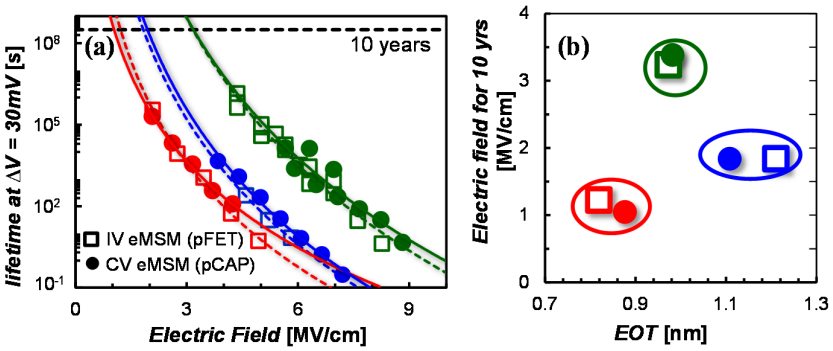



**Fig. 78: (a) Different GF stacks yield virtually identical BTI lifetime extrapolations by *IV* and *CV*-eMSM, (b) regardless of small pMOS/nMOS EOT variations.**

## 3.7 Artefacts of the CV-MSM technique

Performing *C-V*-measurements on UT-EOT devices brings along several difficulties. The major issues are the generation of interface states and the (sometimes stress induced) gate leakage current. Both mechanisms can give

difficulties both for the time-zero device property extraction, as throughout the dynamic $\Delta V_{FB}$ degradation measurement.

### 3.7.1 Effect of interface states

Interface states typically have a two-fold impact on the *C-V*-curve: in high-frequency *C-V* (in the range of 1 MHz), the presence of interface states typically leads to a *stretch-out* and an increase in the conductance signal. As with every *spot-sense* technique, translating the change of the capacitance value towards a $V_{FB}$ shift, is only reliable if there is no *deformation* of the *C-V*-curve. Therefore, comparison of the pre- and post-stress *C-V* characteristic is thought to provide the necessary information on the bulk trapping versus the interface state degradation. The impact of interface traps ($D_{IT}$) on a *C-V* curve is shown in Fig. 79. In both described cases, the same net charging occurs. For Si/SiO2 systems, the density is typically higher close to the band edges.

In high-frequency C-V, the interface states do not follow the AC signal and will be charged according to the DC offset bias. In low-frequency C-V, interface states will charge and discharge along with the AC signal, leading to an additional capacitance and the characteristic '$D_{IT}$ bump'

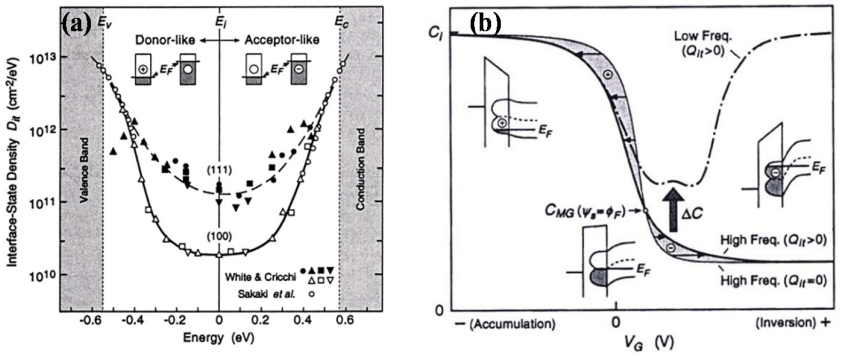



**Fig. 79: (a) The distinction between acceptor- and donor-like interface states is made based on their position w.r.t. the mid-gap. (b) Impact of interface states on the *C-V* curve of a capacitor. [replotted from Schröder06].**

In our test-setups, measurement frequencies of 100 kHz and 1 MHz are used, with the higher frequency lowering the characteristic interface trap bump and lower frequency eliminating series resistance effects. The former is important for correct EOT extraction, while the latter is used during our *C-V*-eMSM test. It was found that increasing the frequency has no intrinsic effect on the stress condition on the device.

Interestingly, in the UT-EOT devices, the $D_{IT}$-bump is sometimes also observed in 1 MHz *C-V* measurements, mostly in RMG stacks. This signifies that some interface traps are able to respond, thus having very low capture and emission times. The fact that a significant $D_{IT}$ bump is also visible at 1 MHz, is believed to be caused by the high overall number of $D_{IT}$, and therefore also the fraction of these very fast traps is increasing.

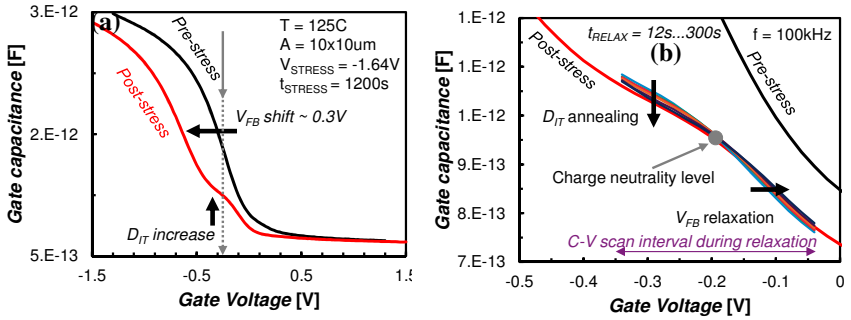



**Fig. 80: (a) Clear increase of interface states ($D_{IT}$) after stressing the capacitor. This will result in seemingly lower $V_{FB}$ shifts with a 'spot-sense' technique. (b) Dynamically tracking a fraction of the *C-V* during relaxation reveals partial annealing of the $D_{IT}$ bump on top of the $V_{FB}$ shift.**

By modifying our 'spot-sense' technique to repeatedly scan a small fraction of the *C-V* around $V_{FB}$, we are continuously scanning the characteristic $D_{IT}$ bump during relaxation. This is shown in Fig. 80. Surprisingly, it was seen that the interface state bump does not only increase, but also decreases during the relaxation (indicated as $D_{IT}$ *annealing* in the figure). A possible explanation for this is that the interface traps that are observed, are actually 'near interface' traps, and behaving accordingly. A similar behavior was

observed by continuous charge-pumping measurements after stress [Grasser11].

The observed $D_{IT}$-bump as described in the above measurements, can have a considerable impact on the apparent $V_{FB}$ shift. As shown in Fig. 81, when increasing stress time (or stress voltages), the *apparent* relaxation can suddenly become a reversed degradation. This can only be due to a decrease of the (near-)interface states.

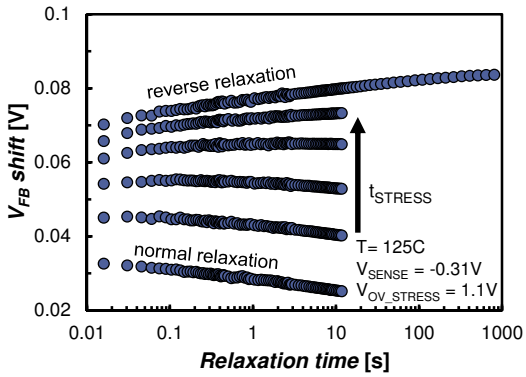



**Fig. 81: With increasing stress time, the apparent $V_{FB}$ tends to degrade more during the relaxation, due to a large increase in (near-)interfacial traps with stress, and subsequent reduction of the $D_{IT}$-bump during relaxation.**

## *3.7.2 Mitigating the interface-state artefacts*

The observation in the previous Section is confirmed in another experiment where we gradually increase the sense voltage condition. Fig. 82 shows the $V_{FB}$ relaxation benchmarked for identical stress times but gradually increasing the sense voltage. As a result, the sensing condition will be higher than the effective $V_{FB}$ and the corresponding $D_{IT}$-bump. As the sense condition diverges from the $D_{IT}$ bump range, the apparent $V_{FB}$ relaxation converges towards the intrinsic $V_{RELAX}$ slope.

Increasing the $V_{SENSE}$ condition has however also the adverse effect on the extracted $\Delta V_{FB}$ and thus lifetime extrapolations, as fewer carriers will be

released from the defect band. This is illustrated by Fig. 73, which shows that by increasing the $V_{SENSE}$, the extracted $\Delta V_{FB}$ will increase (and thus be overestimated). On the other hand, for $V_{SENSE}$ closer to $V_{FB}$, the $\Delta V_{FB}$ at 1ms of relaxation has no physical meaning because of because of the abnormalities in the relaxation curve. Therefore, these data points are withheld in the figure.

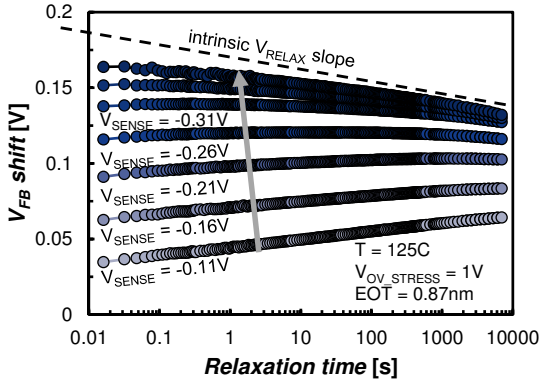



**Fig. 82: Benchmarked for identical stress times, increasing the sense voltage away from the $D_{IT}$-bump helps to recover the effective relaxation slope and thus to find the $V_{FB}$ shift.**

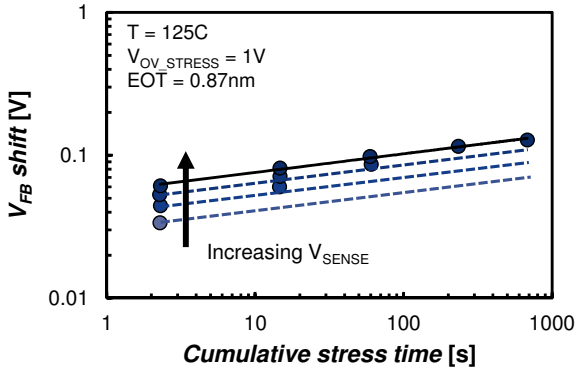



**Fig. 83: The extracted $\Delta V_{FB}$ at 1ms of relaxation for various $V_{SENSE}$ conditions, only the $V_{FB}$ relaxation traces that shown no anomalies are withheld here.**

In a final experiment, depicted in Fig. 84 , the measurement frequency is increased, up to the instrumental limitation of 1 MHz. Even though the anomalous relaxation is slightly mitigated, the effect of the interface states is still prevailing as all the relaxation traces show reverse relaxation, even at the highest available frequency.

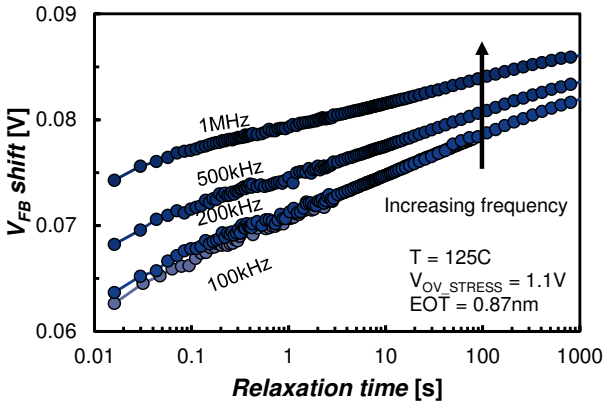



**Fig. 84: By increasing the LCR *C-V*-frequency, the effect of the interface or near-interfae states is lowered, but is still largely prevailing, resulting in anomalous relaxation.**

Concluding, we have to state that there is no straightforward solution to escape from abundant interface state generation and annealing, and its effect on the *C-V* characteristic.

Therefore, in continuation of this work, we use a pragmatic solution: selecting a sensing condition which is systematically higher than the flat-band condition, in the non-$D_{IT}$ affected region. For those $D_{IT}$-affected wafers, the benchmarking will of course be biased with respect to the *IV*-MSM. The presence of $D_{IT}$ can always be verified with a pre- and post-stress measurements.

### *3.7.3 Gate leakage artefacts*

Even though the original purpose of introducing the high-k in the dielectric for ultra-thin EOT was to reduce the gate leakage by increasing the physical

thickness, with downscaled EOT we once again come nearby the direct and Fowler-Nordheim tunneling limits. Moreover, defect generation due to overstress can cause stress-induced leakage current (SILC). This can result in a strongly increased gate leakage after stress, even such that the *C-V*-measurement is dominated by the leakage current. An example is shown in Fig. 85, where the CV becomes unmeasurable after stressing the device.

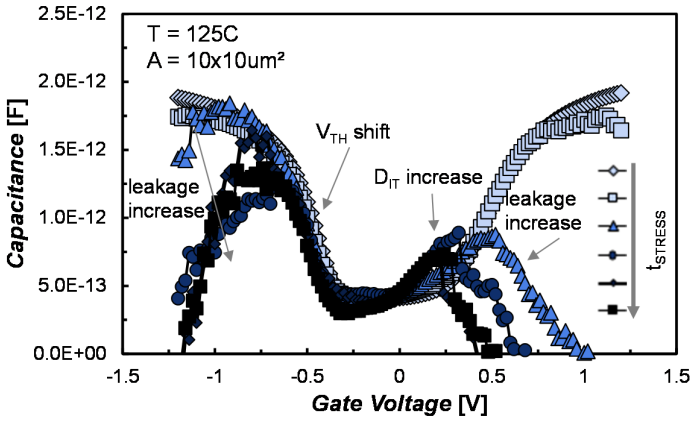



**Fig. 85: Degradation of the *C-V*-curve of an UT-EOT $SiO_2/HfO_2$ device after increasing stress due to Stress Induced Leakage Current (SILC), interface state generation and $V_{TH}$ shift.**

As a result, these extremely leaky or defective devices cannot be taken into account for BTI evaluation. We will however explore and propose alternative methodologies to extract the C-V of these leaky devices in Section 3.11.

## 3.8 Impact of processing on UT-EOT NBTI reliability

In Section 3.6, it was already discussed that utilizing a "gate first"-like integration scheme, minor changes to the process conditions can have a large impact on the BTI lifetime. In this part of the discussion, we look in more detail at potential process knobs that can modulate NBTI in HKMG transistors.

For our assessment method, we will utilize our developed *C-V*-eMSM BTI methodology. Since now plain capacitors can be tested for BTI reliability, quicker feedback for process optimization can be given, due to a significantly lower *turn-around time* (i.e. the time it takes for manufacturing) for capacitor lots.

### *3.8.1 Scavenging in gate-first vs gate-last*

As discussed above, the main approaches for metal gate integration are GF and GL-RMG. The GL approach has become the mainstream integration process because of its broad range of available material options for EWF tuning by allowing low thermal budget after metals deposition, decoupling it from the junction activation by rapid thermal anneal (RTA). Moreover, RMG was shown to considerably enhance the channel stress in short channel devices during dummy gate removal, increasing the benefits from other stress techniques such as embedded-SiGe S/D. However, the smallest EOTs reported to date of the study, had been demonstrated with GF [Ragnarsson09].

The scavenging technique, which was described earlier and which is required to obtain UT-EOT devices, requires a certain thermal budget to activate the diffusion mechanism. Therefore, this scavenging technique has seldom been used in a replacement gate high-k last (RMG HKL) process, mainly due to processing constraints as discussed above. RMG HKL with an Ultra-Thin EOT (UT-EOT) below 7Å has been reported by one source, using doped TiN as a scavenger for the $SiO_2$ interfacial layer [Ando09].

In this case, to obtain scavenging in a GL flow, the scavenger is introduced in the gate metal (TiN) by doping. The Gibbs free energy balance of the metal-oxide will then determine if the interfacial layer is scavenged or regrown (Eq. 3.2). After introducing the scavenger, an 800°C post-metallization anneal (PMA) is subsequently applied (except for the reference stack) to passivate the defects in the high-k stack [Carter03] (Fig. 86).

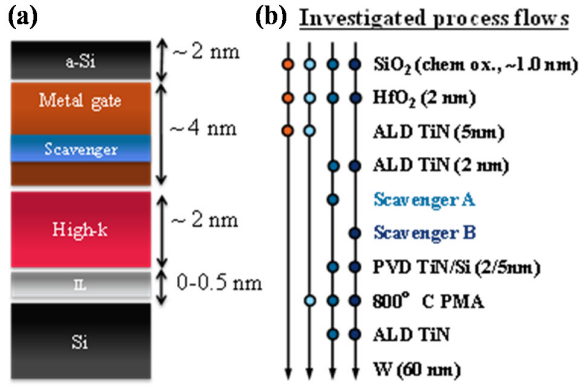



**Fig. 86: (a) Schematic illustration of the gate stack (b) Four process flows are used to investigate the effect of a post-metallization anneal (PMA) and the scavenger on EOT and reliability.**

Comparing the EOTs of the processes, it is clear that the PMA tends to regrow the $SiO_2$ interfacial layer with respect to the reference gate stack. A clear EOT reduction due to interfacial layer scavenging can be observed for both $M_A$ as $M_B$ with respect to the PMA stack without scavenger (Fig. 87).

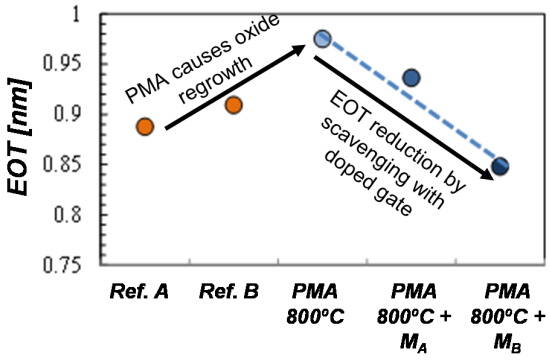



**Fig. 87: EOT extracted from plain *C-V* measurements. The PMA tends to regrow the $SiO_2$ interfacial layer, visible as an increased EOT. A clear EOT reduction can be observed for both scavengers.**

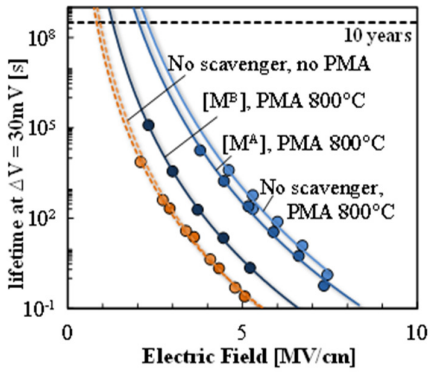



**Fig. 88: *C-V*-eMSM lifetime extrapolations for the RMG GL stacks. The most reliable (no scavenger, PMA 800°C) stack shows a 1.5 MV/cm improvement in maximum allowed electric field over the reference stack.**

As shown in Fig. 88, scavenger $M_B$ improves the tolerable electric field w.r.t. to the reference stacks, whereas the EOT is simultaneously reduced. The PMA thus passivates the defects, thereby yielding an overdrive electric field which is about 1.5 MV/cm higher than the reference stack. Looking at the stack where only the PMA is performed, we observe that is caused a distinct increase the tolerable $E_{OX}$, but at the cost of an increased EOT.

Fig. 89 shows the result of the lifetime benchmark, including earlier obtained data with IV-MSM on GF devices. The yellow arrow indicates the improved reliability over the reference stack while reducing the EOT. However, the steep decrease in GL RMG reliability (indicated with the green dashed line) is still observed about 1.5-2Å earlier than in our most aggressively scaled GF stacks (indicated with the grey line).

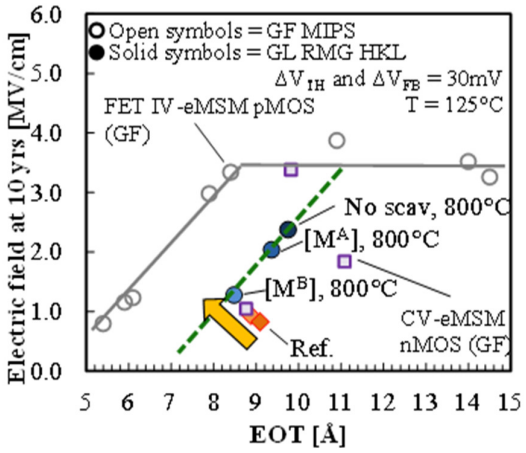



**Fig. 89: The tolerable electric field to achieve 10 years lifetime benchmarking data obtained with IV-eMSM (grey circles) and data obtained with CV-eMSM (solid circles). The colors are corresponding to the process flow in Fig. 86. The yellow arrow indicates the improved reliability over the reference stack while reducing the EOT.**

The offset in the EOT-maximum $E_{OX}$ trends of GF and GL stacks is striking. The major variation in process conditions that can be discerned are the different sequences and gradations of annealing conditions that these deposited gate stacks observe. Whereas in this experiment, the post-metallization anneals where performed up to 800 °C, it was shown later on by Arimura that PMA of 1035°C on GL capacitor devices, could succumb a clear improvement on their reliability, up to the level of typical GF stacks [Arimura13].

Concluding, we have shown that there is no fundamental difference between gate first and gate last gate stack deposition. Both processes suffer from accelerated degradation below a certain $SiO_2$ interfacial layer thickness. Moreover, it was shown that thermal treatment of the gate stack can be a key in improving the reliability.

The open question remains why the accelerated BTI degradation below a certain EOT threshold is observed in Fig. 89. In the next Section, we will study this rapid drop of the reliability below a certain EOT threshold.

## 3.9 Impact of oxide thinning on UT-EOT reliability

There are multiple hypotheses to describe the sudden degradation in BTI lifetime versus EOT. For example, the trend could be ascribed to either a direct effect of the increasing defect density with the $Si/SiO_2$ interlayer scaling down, or as an indirect effect of decreasing the barrier for carriers towards the abundant $HfO_2$ defects. In other words, that the increased accessibility of the high-k stack is contributing to this detrimental trend.

Both the GF and GL results in Fig. 89 show that the reliability rapidly drops for EOTs below a certain threshold, at constant high-k thickness. This view supports the hypothesis that a certain minimal $SiO_2$ thickness is required to obtain sufficient reliability for these gate stacks. We will study this hypothesis in this Section.

According to ab-initio simulations, it was shown that an $SiO_2$ interfacial layer of 0.3-0.5 nm is needed to obtain the full bandgap [Kaneta03] (Fig. 90).

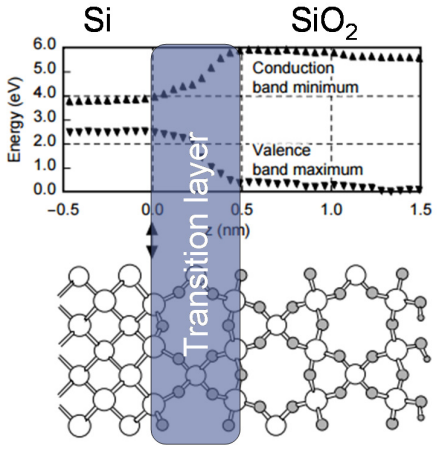



**Fig. 90: The bandgap between Si and the $SiO_2$ interfacial layer increases gradually according to ab-initio simulations. This transition layer has a thickness of 0.3-0.5 nm [replotted from Kaneta03].**

Even minor process variations can give rise to a change in composition and quality of the dielectric stack, an interesting experiment was proposed by

Arimura, in which $Si/SiO_2/HfO_2$ MOS capacitors were formed on varying $SiO_2$ thicknesses. This was obtained by back-etching a thick layer of $SiO_2$. The latter was either grown by rapid thermal oxidation (RTO), or by in-situ steam generation (ISSG), and compared with non-slant-etched chemical oxide [Arimura14]. The amount of back-etching was wafer dependent, as the thick RTO (4.6 nm) and ISSG (5.6 nm) layers, were slant etched in a diluted HF solution with controlled wafer immersion speed to obtain resulting $SiO_2$ thicknesses in the range of 0.4 nm to 1.8 nm *prior* to scavenging.

The tolerable electric field for a 10 years lifetime was subsequently measured with the *C-V*-eMSM measurement technique described in the Sections above. The results from this *slant-etch* experiment are depicted in Fig. 91.

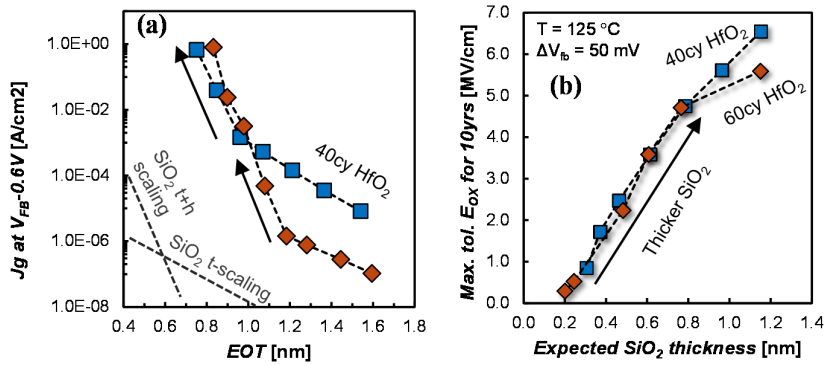



**Fig. 91: (a) Scaling the $SiO_2$ in the slant-edge experiment shows a two exponential slopes for the increase in leakage current. (b) When rescaled to identical interfacial $SiO_2$ thickness, the high-k thickness has no impact on the NBTI reliability. [replotted from Arimura14]**

The two slopes in the *Jg-EOT* curve, indicate another tunneling current mechanism is activated for EOTs in the range of 1nm for 40cy of $HfO_2$. This can only be explained by the $SiO_2$ tunneling barrier *not only decreasing* in thickness, but also *changing in shape*. This corresponds to the ab-initio results shown earlier [Kaneta03], indicating that also the $SiO_2$ barrier *height* will decrease below certain thickness. This experiment thus confirms the minimum $SiO_2$ thickness needed to obtain the full bandgap.

Looking at the NBTI results however, *no bending point* is observed in $SiO_2$ thickness regions between 0.2-0.6nm. Even though it is clear that the $SiO_2$ interfacial layer *thickness* has a direct relation with the NBTI lifetime, regardless of the high-k thickness, the reduced barrier *height* seems not to be influencing the NBTI lifetime.

## 3.10 Understanding the fundamental limits of NBTI-scaling

After analyzing the above experiments, the question still remains what is causing this accelerated NBTI degradation. Even though some hypotheses could already be disproved, a clear mechanism could not be identified yet.

An interesting hint is however given based on the results of the slant-etch experiments, which also revealed the existence of a correlation between a EWF reduction (also called *roll-off*) and enhanced NBTI degradation, shown in Fig. 92. The EWF roll-off with EOT scaling was already widely known before [Shiraishi04, Akasaka06].

The correlation could however only be found on a wafer in which the TiN scavenging MG was removed and a fresh TiN gate was re-deposited after scavenging to restore the EWF. When the scavenging gate was not replaced, the EWF remains low, as there is only *exchange* in charge between the gate metal and the dielectric which results in a dipole, but *not in net charging* (i.e. WF tuning) of the gate stack.

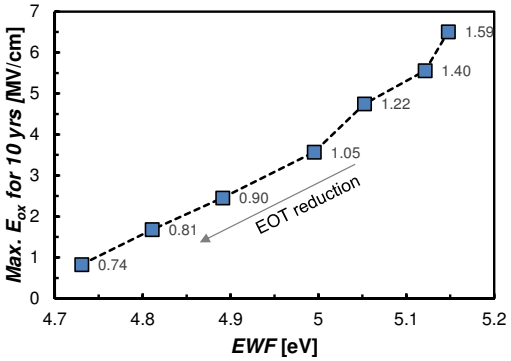



**Fig. 92: Correlation between EWF roll-off and tolerable electric field for 10 years NBTI lifetime measured with *CV*-eMSM [based on Arimura14].**

Bersuker *et al.,* proposed a mechanism that suggests that the roll-off phenomenon is caused by enhanced positive-charge generation within the interfacial $SiO_2$ layer, when its thickness falls below a certain critical value [Bersuker10]. The mechanism, based on Fermi-level pinning via oxygen vacancies, is illustrated in Fig. 93.

The generation of oxygen vacancies is expected to be significantly higher when oxygen species are consumed from the *transitional* $SiO_x$ layer adjacent to the Si substrate. This will result in a static effect, i.e. the EWF-roll off by as the EOT is decreased, because the $SiO_2$ layer is decreasing. At the same time, the oxygen vacancy location comes closer to the Si-$SiO_2$ interface, thereby enhancing the accessibility from the channel, i.e. leading to more charge trapping and de-trapping.

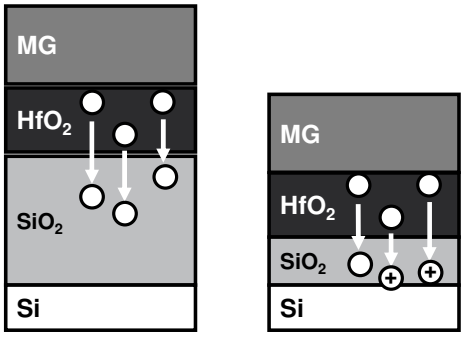



**Fig. 93: The mechanism for EWF roll-off as proposed by Bersuker. The arrows illustrate the vacancy diffusion by the movement of oxygen atoms in the opposite direction. [replotted from Bersuker10]**

Another mechanism was proposed by, Arimura *et al.*. He pointed out that the fixed charges (again referring to the EWF roll-off) near the $SiO_2$/high-k interface could also change the defect band alignment of the high-k layer [Arimura14]. As a result, *the remaining fixed charges modify the alignment of the high-k defects*, in similar way as rare-earth materials were used to shift the defect levels up for PBTI [Kaczer09].

A possible way to distinguish between both mechanisms would be by assessing the (near-)interface state density by means of charge pumping or defect spectroscopy (TSCIS). In the case of the former mechanism, the interface or near-interface state density is expected to be correlated to EOT reduction and EWF roll-off. In the case of the latter mechanism, indirect impact of the fixed charges on the defect level, the (near-)interface state density is not expected to be directly correlated to the EWF or EOT.

Due to the fact that this is a capacitor-only wafer, these charge-pumping or TSCIS experiments could not be done. However, as discussed earlier in this Chapter, Cho *et al.* showed that for UT-EOT devices, the main degradation component was due to bulk trapping [Cho11], thereby making the option of Arimura's indirect effect due to defect band alignment by fixed oxide charges more plausible.

# 3.11 Alternative methods for C-V extraction

Some devices have such a high defectivity already from their fabrication. Even though their reliability might already be jeopardized because of this, it is interesting to be able to at least extract the EOT, for example to estimate the remaining $SiO_2$ interfacial layer thickness. For this reason, we develope a method based on a pulsed-measurement, where the displacement current of the device is measured directly.

Another issue is that nanoscale devices fundamentally differ from large area devices, for example due to changes in ratios of overlap and sidewall effects with respect to the bulk of the device. The signal-to-noise ratio of a C-V measurement is however directly related to the area of the device, which makes nanoscale devices unmeasurable. To extract the *C-V* of nanoscale devices, we present an on-chip characterization methodologies based on the charge-based-capacitance technique.

Both methodologies will be described in Section 3.11.1 and Section 3.11.2.

## *3.11.1 Single-Pulse CV*

In this Section, the single-pulse *C-V* measurement approach is calibrated, verified and utilized to capture the *C-V* characteristics of leaky devices. The accuracy and reliability of this testing method is subsequently examined and verified. A similar pulsed-*C-V* method was earlier proposed by Ji *et al.* for mobility extraction [Ji13].

An arbitrary waveform generator (Keithley 4200PIV system) was implemented to generate a voltage pulse waveform. By applying a triangular pulse with a constant ramp rate $\Delta V_g(t)/\Delta t$ on a capacitor, a displacement current $i_{dis}$ will flow, proportional to the (local) capacitance $C$:

$$i_{dis}(t) = C(V_g)\frac{\partial V_g(t)}{\partial t} \quad . \tag{3.8}$$

Initially, we calibrate our system with commercially available ceramic capacitors, which have similar values as typical gate capacitances for 10x10 $um^2$ devices. The resulting displacement current is then tuned by varying the

ramp rate of the triangular pulse, such that the displacement current is maximized within the used measurement range of the system. As such, the noise level of the measurements is minimized. The applied triangular pulse is shown in Fig. 94. As can be seen, the observed displacement currents are constant during the rising and falling edges of the applied pulse.

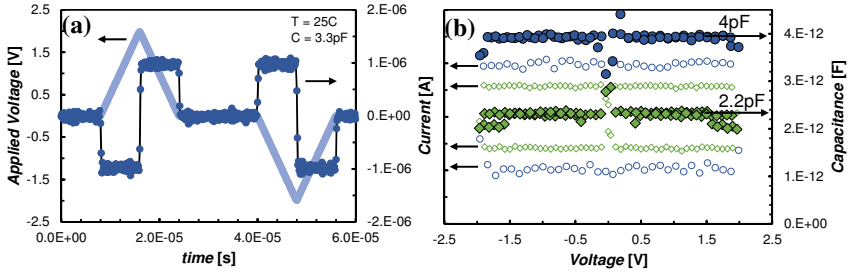



**Fig. 94: (a) Triangular pulse shape generated by arbitrary waveform generator and resulting displacement current. (b) Calibration of the setup with ceramic capacitors, yielding a constant C-V curve at the respective calibration capacitances of 4 and 2.2 pF, both for negative as positive pulses.**

Subsequently, this setup can be used to measure MOSFET capacitances. The pulses are applied to the gate, whereas the source/drain and bulk displacement currents are measured separately. As such, not only the overall capacitance, but also the gate-to-channel ($C_{GC}$), gate-to-bulk ($C_{GB}$) and gate-to-all capacitances ($C_{GA}$) can be derived:

$$C_{GC} = \frac{I_S + I_D}{\partial V_g / \partial t} \quad , \tag{3.9}$$

$$C_{GB} = \frac{I_B}{\partial V_g / \partial t} \quad , \tag{3.10}$$

$$C_{GA} = C_{GB} + C_{GC} = \frac{I_B + I_S + I_D}{\partial V_g / \partial t} \quad . \tag{3.11}$$

The displacement currents measured at the source/drain junctions for charging and discharging a 10x10 µm² high-k/MG pFET in inversion are shown in Fig. 95. Both the up as the down sweep are considered and overlaid, since artificial delays in the response current time could potentially result in a shifted *C-V* curve. The fully derived capacitance-voltage characteristics for this transistor in inversion and accumulation are shown in Fig. 96. The resulting EOT of this transistor is a conservative 11.5 Å.

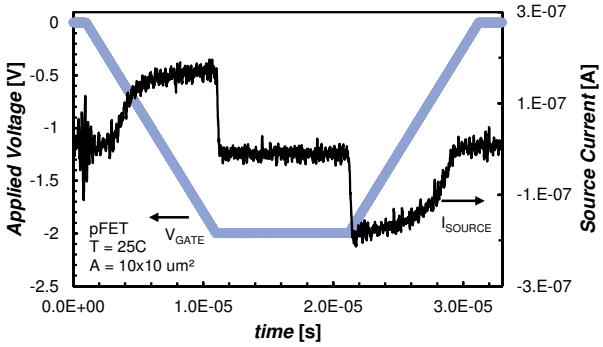



**Fig. 95: Applied trapezoidal voltage and resulting displacement current at the source of 10x10 um² Si/$SiO_2$/$HfO_2$ transistor, eventually yielding the $C_{GB}$ (partial from 0 to -2V).**

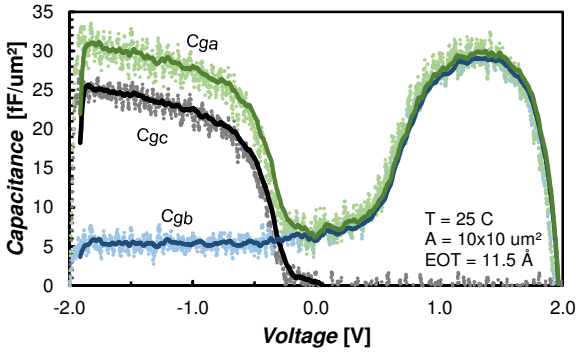



**Fig. 96: Conversion of the directly measured displacement currents into the $C_{GB}$, $C_{GC}$ and $C_{GA}$-voltage characteristics for a 10x10 um² pFET with thin-EOT.**

It should be noted that when this technique is used, the effect of the *geometric component* should be taken into account, for example: when the selected device has a channel length of 100 µm, the generated charges will remain in the channel for several µs [Bosch93]. This is because the flow of the carriers is determined (at long times) by a concentration gradient-induced diffusion component towards the source/drain, as described with the equation below:

$$N(t) = \frac{8N(0)}{\pi^2} e^{\left(-\frac{D_N \pi^2}{L_G^2} \cdot t\right)}, \qquad (3.12)$$

where $L_G$ is the gate length and $N(0)$ the initial sheet of charge density in the channel and $D_N$ the diffusion constant of the channel carriers.

The channel length consideration is similar as to be made with typical charge-pumping techniques [Groeseneken84]. Also here, it will impact the observed displacement currents, which are time-critical as discussed above. However, if the channel length is 10 µm or below, the vast majority of inversion charges in the channel can flow back to the source drain regions within 100 ns, which is shorter than the applied pulse time.

This technique proves being a viable alternative for typical *C-V* measurements, even on (leaky) thin-EOT devices. However, this technique only works on large-area devices, as the noise of the setup is larger than the observed signal in small devices.

### *3.11.2 On-chip charge-based capacitance measurements*

In this Section, an on-chip characterization circuit for *nanoscale* MOSFET *C-V* analysis is presented. Capacitance measurements using a quasi-static charged-based measurement technique (CBCM) with atto-Farad resolution is shown.

The experimental characterization of the gate oxide in nanoscale FinFET (and eventually nanowire) devices is impeded due to the difficulty in measuring the capacitance that is typically *below* the femtofarad scale. The multi-dimensional nature of these devices makes it impossible to have

representative large-area device for channel charge characterization, as is the case with bulk CMOS technology.

In typical capacitance measurements, the signal-to-noise ratio is proportional to the small-signal frequency. Due to instrumentation limitations (effect of series resistance, parasitic capacitances, …) the frequency cannot be ever-increased off-chip. Therefore, capacitance characterization relies on experimental data obtained from a large number of devices connected in parallel. Increasing the total device capacitance to measurable quantities masks the inherent device-to-device variability at the nanoscale. This makes experimental characterization of gate capacitance in nanoscale devices challenging. A solution to this issue is the charge-based-capacitance measurements (CBCM).

In CBCM, the device's capacitance is measured by separating and accumulating the charges required to charge and discharge the device. The accumulated flow of charging and discharging currents, generated by supplying a GHz-frequency pulse at the gate of the transistor, can be measured as a continuous current with a static measurement, i.e. with a SMU. This simple principle is described in literature, and typically also used to assess parasitic interconnect capacitances [McGaughy97].

By providing non-overlapping pulses to the pass transistors controlling the charging (pull-up) and discharging (pull-down) branches, the DUT's gate is charged and discharged sequentially. The effects of parasitic leakage, either in the DUT or via the pass transistors cannot be extracted by means of a simple differential measurement. Therefore, the branches for charging and discharging the DUT are designed in twofold. The complementary (dis)charging branches only serve to measure the parasitic currents, without effectively altering the charge on the DUT's gate. As such, the parasitic currents can be subtracted from the charging currents. The proposed circuit and an illustration of the resulting currents are illustrated in Fig. 97 and Fig. 98.

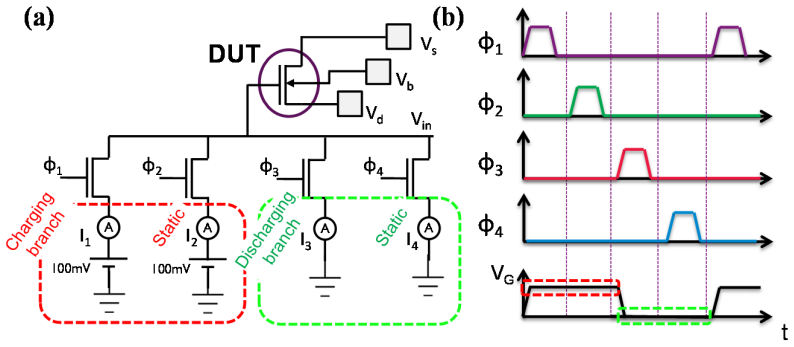



**Fig. 97: (a) Principle of CBCM measurement on a device. The net current flowing through each of the feeding transistors represents the charging, static leakage at high $V_G$, discharging and static leakage at low $V_G$ currents respectively. (b) Each of the branches is sequentially opened by transmitting non-overlapping pulses to its controlling transistors.**

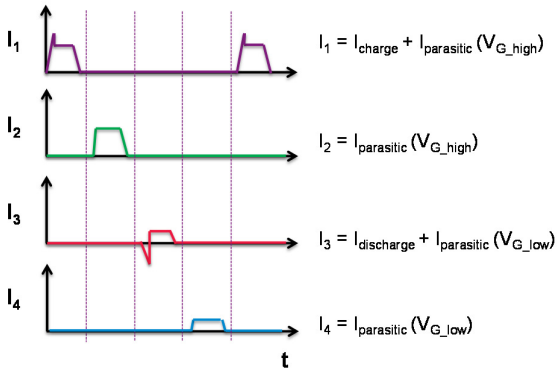



**Fig. 98: The resulting currents in each of the branches of the device. The complementary branches with currents $I_2$ and $I_4$ will only accumulate parasitic currents.**

The differential charging current can be described as follows:

$$I_{CBCM} = (I_1 - I_2) = -(I_3 - I_4) \tag{3.13}$$

The average current $I_{CBCM}$ the represents the charges necessary to fully charge (or discharge) the capacitance of the DUT ($C_{DUT}$) at a fixed small signal range of $V_{AC}$ at the frequency $f$. The resulting capacitance can then be calculated as:

$$C_{DUT} = I_{CBCM} \Big/ f.V_{AC} \quad , \tag{3.14}$$

and can be evaluated at any bias condition can by controlling the separate source, drain and bulk connections.

One of the key components of this methodology is that the pull-up and pull-down switches are driven by non-overlapping periodic signals in order to avoid short-circuit currents that may perturb the charging current $I_{CBCM}$. This means that the non-overlapping pulses also need to be generated on-chip. This can be established by connecting simple combinatory logic on a the output of a ring-oscillator (Fig. 99).

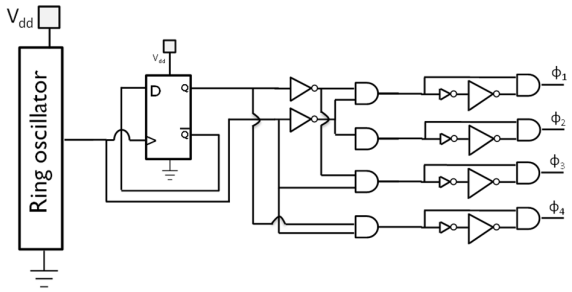



**Fig. 99: Simple combinatory logic, utilizing a D-flipflop and a multiplexer and a delay chain converts the output of a ring-oscillator into four non-overlapping pulses.**

The resolution limit of this type of CBCM will be caused by (1) the mismatch of the parasitic capacitances between the sets pull-up/pull-down branches and (2) by mismatches in the pulse duration generated by the combinatory logic. By upsizing the transistors in both the combinatory chain as in the pull-up/pull-down branches, the error can be minimized.

In the picture below, simulations of the entire circuit are shown in a HSPICE 16nm predictive technology model (PTM). The working principle is clear. It is shown that with a RO frequency of ~ 8GHz, DC currents of 300 nA are outputted, resulting in a capacitance of 0.15fF, i.e. the gate capacitance of a 48x16nm (*W*x*L*) MOSFET at a bias condition of 1V (Fig. 100). This is a measurable quantity with any off-the-shelf SMU.

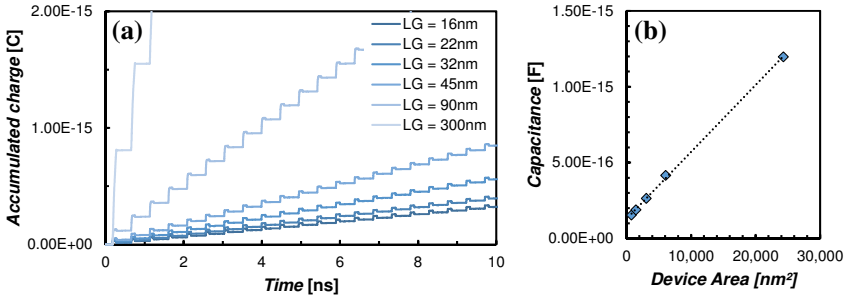



**Fig. 100: (a) Simulated accumulated charge via pull-up line for various DUT sizes, resulting in a current level of 300nA for the smallest device and (b) corresponding capacitances as would-be extracted with DC SMU.**

Fig. 101 shows the layout of the CBCM structure that was developed in imec's first 14nm mask set. Up to date, no silicon devices in this mask set are yet available.

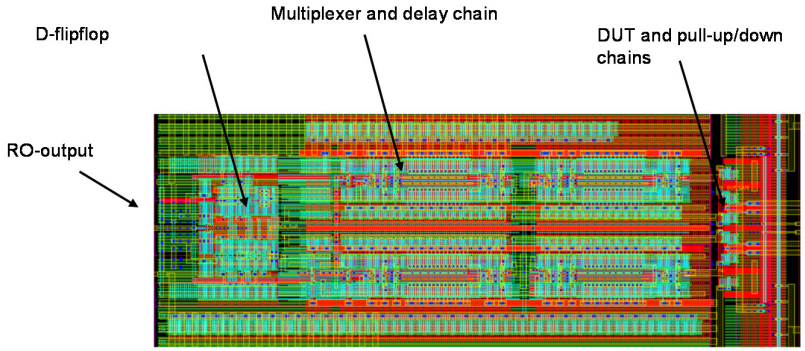



**Fig. 101: Corresponding layout of the proposed CBCM in imec's 14nm technology.**

## 3.12 Conclusions

In this Chapter, measurement techniques were developed allowing systematic assessment of UT-EOT devices, both in their initial characteristics, by single-pulse *C-V* method, as for their long term degradation by developing a *CV*-eMSM technique. The latter was found to be very effective in systematic screening of experimental gate stacks. In the margin of this technique, a single pulse *C-V*-technique was presented in which the *C-V* characteristics of leaky devices can be extracted. Also, a CBCM circuit was presented, capable of extracting device capacitances of nanoscale devices.

Using the *CV*-eMSM technique and in the pursuit of improving the NBTI lifetime, it was found that gate stack annealing conditions play an important role in determining gate stack quality, and appeared to be the major difference between gate first and gate last device stacks. It was shown that by applying the right annealing conditions for gate last stacks, they can get up to the reliability of their gate first counterparts, albeit still largely under the ITRS roadmap specification.

The fundamental origin of the accelerated BTI trend appears to be related to the scavenging of the $SiO_2$ interfacial layer, causing oxygen vacancies near the $SiO_2/HfO_2$ interface, and which get charged by holes charges in the high-*k*. The experimental observations of the correlation between EWF roll-off and BTI lifetime, provides a strong indication that either scavenging related-vacancies are brought closer to the oxide when the interfacial SiO2 is thinned, or that the fixed charges can modify the alignment of the $HfO_2$ defect band-k, such that the high-k defect level becomes *more* accessible when the EOT is scaled.

In order to improve BTI reliability for gate stacks with an EOT below 1nm, more fundamental solutions will have to be found, most probably no longer relying on oxygen scavenging. In the meanwhile, it appears that the entire semiconductor industry is struggling with this issue, as at the moment of writing this thesis, dielectric thicknesses have barely been scaled below 1nm of EOT. The short-term solution for industry to maintain gate control for reduced channel lengths, has been to produce MuG- or FinFET devices, whereas in the long-term, further EOT scaling can still be beneficial, not only

for improved performance, but also for reducing variability. Therefore, the continued study for BTI on UT-EOT remains a pertinent research domain. Moreover, the above-presented techniques can, for example, also be used for assessing gate-stack quality on novel (high mobility) channel materials, similar as those introduced in Chapter 6.

Finally, we provided methodologies, single-pulse CB and on-chip CBCM to evaluate the CV on leaky and nanoscale devices, respectively.

# 3.13 References

[Akasaka06] Akasaka. Y. et al., "Modified oxygen vacancy induced Fermi level pinning model extendable to P-metal pinning", Japanese Journal of Applied Physics, 45, pp. 1289, (2006).

[Ando09] Ando T. et al., "Understanding Mobility Mechanisms in Extremely Scaled HfO2 (EOT 0.42 nm) Using Remote Interfacial Layer Scavenging Technique and Vt-tuning Dipoles with Gate-First Process", in Proc. International Electron Devices Meeting, pp. 423-426, (2009).

[Ando11] Ando T. et al., "Origin of Effective Work Function Roll-off Behavior for Replacement Gate Process Studied by Low-temperature Interfacial Layer Scavenging Technique", as discussed at IEEE Semiconductor Interface Specialist Conference (SISC), Arlington, VA, Dec. 1-3, 2011.

[Arimura14] Arimura H. et al, "Guidelines for reducing NBTI based on its correlation with effective work function studied by CV-BTI on high-k first MOS capacitors with slant etched SiO2", in Proc. IRPS,

[Baklanov07] Baklanov M., Maex K, Green M., "Dielectric Films for Advanced Microelectronics", Wiley, (2007).

[Bersuker10] Bersuker G. et al.," Origin of the Flatband-Voltage Roll-Off Phenomenon in Metal/High-k Gate Stacks", in Trans. On Electron Devices, Vol. 57, pp. 2047, (2010).

[Bosch93] G. V. D. Bosch, G. Groeseneken and H. E. Maes, "On the geometric component of charge-pumping current in MOSFET's", IEEE Electron Dev. Lett., vol. 14, no. 3, pp. 107-109, (1993).

[Carter03] "Passivation and interface state density of SiO2/HfO2-based/polycrystalline-Si gate stacks," in Appl. Phys. Lett. 83, 533, (2003).

[Cho11] Cho M. et al., "Interface Trap Characterization of a 5.8-Å EOT p-MOSFET Using High-Frequency On-Chip Ring Oscillator Charge Pumping Technique", in IEEE Trans. Elec. Devices, Vol. 58, No. 10, pp. 3342-3349, (2011).

[Choi09] Choi K. et al., "Extremely Scaled Gate-First High-k/Metal Gate Stack with EOT of 0.55 nm Using Novel Interfacial Layer Scavenging Techniques for 22nm Technology Node and Beyond", in VLSI Tech. Digest, pp. 138-139, (2009).

[Choi10] Choi C. and. Lee J. C, "Scaling equivalent oxide thickness with flat band voltage (VFB) modulation using in situ Ti and Hf interposed in a metal/high-k gate stack", J. Appl. Phys. Vol. 108, pp. 64107 (2010).

[Franco10] Franco J. et al, "Improvements of NBTI reliability in SiGe p-FETs", in Proc. of International Reliability Physics Symposium (IRPS), pp. 1082-1085, (2010).

[Franco12] Franco J. et al., "Impact of single charged gate oxide defects on the performance and scaling of nanoscaled FETs", in Proc. . IEDM, pp. 5A.4.1 - 5A.4.6, (2012).

[Franco14] Franco J. et al., "Understanding the suppressed charge trapping in relaxed- and strained-Ge/$SiO_2$/$HfO_2$ pMOSFETs and implications for the screening of alternative high-mobility substrate/dielectric CMOS gate stacks", in Proc. IEDM, pp. 397-400, (2013).

[Grasser11] Grasser T. et al.," The 'permanent' component of NBTI: Composition and annealing", in Proc. IRPS, pp.6A.2.1-6A2.9, (2011).

[Groeseneken10] Groeseneken G. et al., "Trends and perspectives for electrical characterization and reliability assessment in advanced CMOS technologies", in Proc. ESSDERC, pp. 64-72, (2010).

[Groeseneken84] G. Groeseneken, H. E. Maes, N. Beltran, and R. F. De Keersmaecker, "A reliable approach to charge-pumping measurements in MOS transistors," IEEE Trans. Electron Devices, vol. ED-31, p. 42, (1984).

[Hauser98] Hauser J.R. and Ahmed K., "Characterization of Ultra-Thin Oxides Using Electrical C-V and *I-V* Measurements", in Characterization Metrology ULSI Technology, 449, pp. 235-39, (1998).

[Ho12] Ho T.J.J. et al., "Are Interface State Generation and Positive Oxide Charge Trapping Under Negative-Bias Temperature Stressing Correlated or Coupled?" in IEEE Trans. on Elec. Devices, Vol 59, No. 4, pp. 1013-1022, (2012).

[Houssa06] Houssa M. et al, "Electrical properties of high-k gate dielectrics: Challenges, current issues, and possible solutions", in Material Science and Engineering, Vol. 51, pp. 37-85, (2006).

[Huang09] Huang J. et al., "Gate First High-k/Metal Gate Stacks with Zero SiOx Interface Achieving EOT=0.59nm for 16nm Application", in VLSI Tech. Digest, pp. 34-35, 2009.

[Huard03] Huard V., Monsieur F., Ribes G, and Bruyere S., "Evidence for hydrogen-related defects during NBTI stress in p-MOSFETs", in Proc. IRPS, pp. 178, (2003).

[Jayachandran15] Jayachandran S., et al, "Deposition of O atomic layers on Si(100) substrates for epitaxial Si-Osuperlattices: investigation of the surface chemistry", in Applied Surface Science 324, pp. 251–257, (2015).

[Ji13] Ji Z. et al., "A new Ultra-Fast Single Pulse technique (UFSP) for channel effective mobility evaluation in MOSFETs", in Proc. IEEE Int. Conf. Microelectronic Test Structures, pp. 64-68, (2013).

[Kaczer09] Kaczer B., Veloso A., Aoulaiche M. and Groeseneken G., "Significant reduction of positive bias temperature instability in high-k/metal-gate nFETs by incorporation of rare earth metals", in Microelectronic Engineering, vol. 86, no. 7-9, pp. 1894-1896, (2009).

[Kaneta03] Kaneta C., Yamasaki T and Kosaka Y, "Nano-Scale Simulation for Advanced Gate Dielectrics", in Fujitsu Sci. Tech. Journal, Vol. 39, pp. 106-118, (2003).

[Kerber09] Kerber A. et al., "Reliability Challenges for CMOS Technology Qualifications With Hafnium Oxide/Titanium Nitride Gate Stacks", in IEEE Trans. on Device and Materials Reliability, vol. 9, is. 2, pp. 147-162, (2009).

[Kim04] Kim H. et al., "Engineering chemically abrupt high-k metal oxide/silicon interfaces using an oxygen-gettering metal overlayer", in Journal of Applied Physics. vol. 96, No. 6, pp. 3467-3472, (2004).

[Kim04] Kim H. et al., "Engineering chemically abrupt high-k metal oxide ⁄ silicon interfaces using an oxygen-gettering metal overlayer" J. Appl. Phys. Vol. 96, pp. 3467 (2004).

[McGaughy97] McGaughy et al., "A Simple Method for On-Chip, Sub-Femto Farad Interconnect Capacitance Measurement", in IEEE Trans. Elec. Dev., vol. 18, no. 1, (1997).

[Nicollian82] Nicollian E.H. and Brews J. R., "Metal Oxide Silicon Capacitor at Intermediate and High Frequencies", in MOS (Metal Oxide Semiconductor) Physics and Technology, pp. 99-175, (1982).

[Ragnarsson09] Ragnarsson L.-Å. et al., "Ultra low-EOT (5Å) gate-first and gate-last high performance CMOS achieved by gate-electrode optimization", in Proc. IEDM, pp. 663-666, (2009).

[Ricco88] Ricco B. et al., "Oxide-thickness determination in thin-insulator MOS structures", in IEEE Trans. Elec. Dev., vol. 35, pp. 432-438, (1988).

[Schröder06] Schröder D.K., "Semiconductor Material and Device Characterization", 3rd Edition, Wiley, (2006).

[Shiraishi04] Shiraishi K. et al, "Physics in Fermi level pinning at the polySi/Hf-based high-k oxide interface", in VLSI Tech. Symp., pp. 43, (2004).

[Spessot14] Spessot A. et al., "Impact of Off State Stress on advanced high-K metal gate NMOSFETs", in Proc. ESSDERC, pp. 365-368, (2014).

[Toledano11] Toledano M. et al., "Fast VTH Transients After the Program/Erase of Flash Memory Stacks With High- Dielectrics", in IEEE Transactions on Electron Devices, pp. 631-640, (2011).

# Chapter 4: Unifying RTN, BTI and SILC in nanoscale devices

We show how RTN, BTI and SILC are correlated and how the observed effects can be explained with a *refined* 4-state non-radiative multiphonon model (NMP). By studying single trap activated leakage paths, it is shown that additional gate tunneling current in nanoscale FETs can be ascribed to thermally activated defect states.

## 4.1 Introduction

The understanding of oxide trap behavior is crucial for a number of device reliability issues such as Bias Temperature Instabilities (BTI), Random Telegraph Noise (RTN), Time-Dependent Dielectric Breakdown (TDDB) and Hot Carrier Degradation (HCD). Moreover, as a consequence of scaling the device dimensions, degradation and fluctuations in drain and gate leakage currents become more pronounced, even in a way that they could seriously affect device performance. This increased impact of single traps however, also allows us to study the impact of single trapping phenomena. Examining these phenomena in nm-sized FETs can give insight in the underlying physical principles.

In this Chapter, we will first study the correlations of the typical macroscopic versus microscopic degradation effects of BTI and RTN. Subsequently, based on a new set of measurements on nanoscale devices, we investigate how both these degradation phenomena are related to stress induced leakage currents (SILC) in a *microscopic* way. Finally, we will show how the observed phenomena can be explained with a *refined* 4-state non-radiative multiphonon model [Grasser10a].

## 4.2 Macroscopic versus microscopic behavior of BTI and RTN

In this Section, we briefly look at how BTI degradation and low-frequency noise or RTN are observed, both from the traditional "top-down" approach (deducing the microscopic mechanisms of average BTI degradation in large devices) as by the recently introduced "bottom-up" approach in nanoscale devices, in which nanoscale BTI is understood in terms of charging and discharging of individual defects.

### *4.2.1 BTI degradation*

As discussed in Chapter 2, it is well established that in small devices, charge trapping and de-trapping of single defects can significantly alter the channel current, as shown in Franco et al.'s "*ultimate*" BTI experiment [Franco12]. In a large device, after stressing the device, the threshold voltage will show a quasi-continuous relaxation whereas in a nanoscale device, discrete steps in the $V_{TH}$ are observed. This is depicted in Fig. 102.

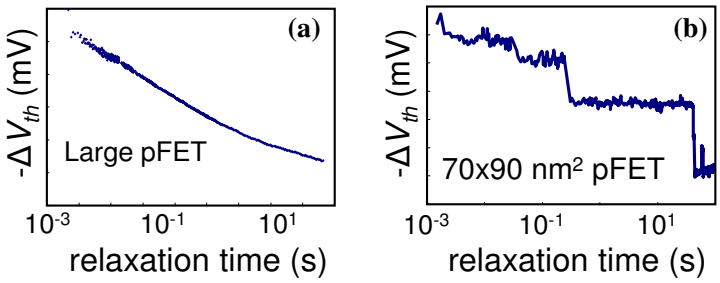



**Fig. 102: Phenomenologically, BTI exposes itself as (a) ~log (t) relaxation in large-area devices and (b) shows a step-wise relaxation in nanoscale devices [replotted from Kaczer08].**

In the observed NBTI mechanisms the defect states exhibit a wide distribution of time scales [Kaczer08]. Moreover, every nanoscaled device is expected to behave differently due to another configuration of the dopants in the channel and of the physical location and configuration of the oxide defects.

We emphasize again that the significant height variations of the steps cannot be simply ascribed to defect depth in the oxide. It is these specific step heights and relaxation times that will help us to build a capture and emission time (CET) maps, based on the TDDS (time-dependent-defect-spectroscopy) technique, discussed in [Grasser10b]. These CET maps show some interesting bias and temperature dependencies which will allow to establish a link between RTN and BTI, as will be discussed further.

### *4.2.2 Random-telegraph noise*

It was reported that both 1/f noise and NBTI relaxation are due to defects with very similar properties [Kaczer09]. Gate-referred noise spectra measured on large devices showed clear 1/f dependence (Fig. 103(a)), which can be explained by a superposition of states with widely distributed time scales. In small devices, noise typically manifests itself as a random telegraph signal (Fig. 105(b)). Irrespective of the existence of a common physical mechanism, the phenomenological similarities were evident.

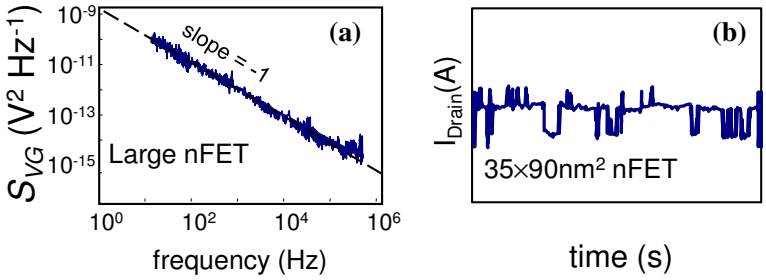



**Fig. 103: Both (a) 1/f noise measured on a large-area device as (b) random telegraph noise on a nanoscale device are originating from trapping and de-trapping of charges in the gate oxide [replotted from Kaczer09 and Toledano12].**

The first modeling efforts of RTN and 1/f noise dates back to the work of McWhorter [McWhorter57]. Kirton and Uren have used a lattice relaxation multiphonon emission (LRME) process [Kirton89]. Grasser *et al.*, however, were the first group to successfully describe the observed voltage and temperature dependences of these RTN states using a NMP model

[Grasser09]. Consequently, he concluded that the defects responsible for this random telegraph noise (the same defects, as described above, which have been suspected to also be the origin blocks of 1/f noise), could also play a role in NBTI.

In the next Section, we explain how "switching traps" in this NMP model, that was already suggested for the recoverable component of NBTI, can also accurately describe the bias and temperature dependence of RTN.

### *4.2.3 Explaining BTI and RTN with 4-state model*

Fig. 104 shows the defect model, as proposed by Grasser, including two stable and two metastable states [Grasser10a]. Fig. 104 shows how defects can be responsible for both RTN and NBTI. Each defect has two stable states, 1 and 2, and possible two metastable states 1' and 2'. Even though all the states are present, the energetic position of each state, based on the particular configuration of the defect, will determine how the trap behaves, i.e. like a fixed positive or a switching trap.

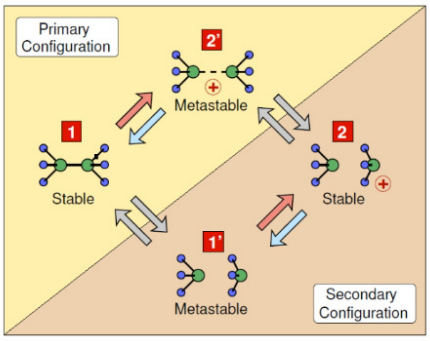



**Fig. 104: Four state model for oxide defects as proposed by [Grasser10b] based on TDDS experiments. [replotted from Grasser10a].**

A qualitative explanation for the model can be given as follows: prior to stress, the defects are in quasi-equilibrium and depending on their respective capture ($\tau_c$) and emission times ($\tau_e$), they are either neutral (state 1 or 1'), positively (state 2 or 2) or negatively charged (not illustrated here). A fraction of the defects, however, randomly capture and release charge, e.g. by

continuously switching back and forth from *neutral* state 1 to the *positively charged* metastable state 2', thereby creating the (visible) RTN signal. When the bias is changed to the stress voltage as in a typical NBTI experiment, this equilibrium is disrupted by the strong bias dependence of the capture and emission time constants, as depicted in Fig. 105. Depending on the combination of capture and emission time constants of the defect at this stress voltage, a previously neutral defect can either stay neutral (when $\tau_c$ is still smaller than $\tau_e$), become charged (when $\tau_e > \tau_c$) or start to produce RTN (when $\tau_c \sim \tau_e$). When the bias is switched back to its initial value, each defect will respond according to its proper emission time constant, after which the system will converge to the previous equilibrium, typically by releasing charges, visible as the discrete steps in the nanoscale devices. As an example, Fig. 105 shows that charge emission can be drastically accelerated for some defects by applying a gate bias into the depletion regime. Finally, it should be noted that some defects do not exhibit the above described bias dependence either for their $\tau_c$ as their $\tau_e$, making it *permanently switching traps*.

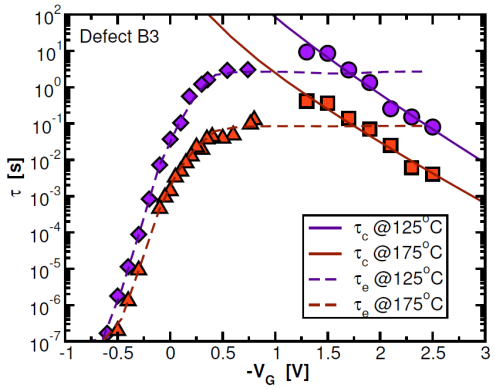



**Fig. 105: Extracted capture ($\tau_c$) and emission time constants ($\tau_e$) for a defects at a various temperatures: both time constants ($\tau_c$) of defect 'B3' are bias-dependent [replotted from Grasser14].**

Grasser *et al.* showed that the *nuances* of the frequency dependent degradation (i.e. the result of AC-stress) could only be explained by the existence of *metastable* states, which yields a *frequency-dependence of the capture time* of individual defects [Grasser12]. In a capture/emission process

from the first order (i.e. having only two states), this specific frequency-dependence would not exist. The latter could only be explained with the above described 4-state model which thus also includes metastable states. This is depicted in Fig. 106 in which the occupancy of state 2 is determined by its *effective capture time*, which is determined by the product of the occupancy of metastable state 2' and the time for passing from metastable state 2' towards stable state 2 ($\tau_{2'2}$). Initially, at low frequencies state 2' from Fig. 104 can always be charged because $\tau_{c12'} > \tau_{e2'1}$ due to the bias conditions. When the frequency is increased, both $\tau_{c12'} > \tau_{e2'1}$ become larger than the inverse frequency. As a result state 2' is most likely discharged in every cycle, thus also inhibiting the transition from metastable state 2' towards stable state 2. Thus for higher frequencies, the metastable state 2' acts like a low-pass filter. In a two-state model only the final occupancy level would depend on frequency but not the *effective capture time.* Experimentally, this translates in a *much longer accumulated stress time* before state 2 will become occupied.

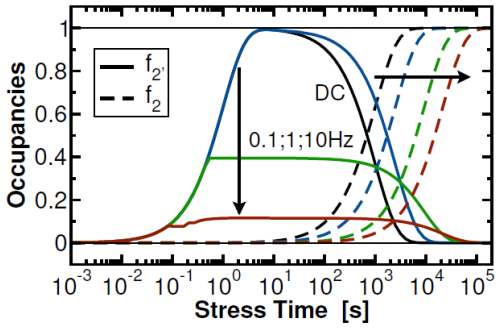



**Fig. 106: The description of the frequency-dependent charging and discharging of state 2 via a metastable state 2', represented as the occupancies [replotted from Grasser12].**

## 4.3 Stress-induced leakage current

In Section 4.2 we explained the link between BTI and RTN via the NMP model by the properties of individual defects and their impact on $I_D$. Here, we discuss the link of Stress-Induced Leakage Current (SILC) with oxide traps

by macroscopic observations (i.e. experimental observations on large devices).

### 4.3.1 Macroscopic observations

Stressing the device results in generation of additional defects, which lead to a distinctive increase of the gate leakage current. On thick-oxide, large-area devices, the correlation between increased gate leakage current and charge trap density has already been shown in the past [DeBlauwe96],[Crupi04]. In *ultra-thin oxide* devices, i.e. physical thicknesses below 2nm, the oxide thickness is sufficiently reduced so the trap generation rate decreases by many orders of magnitude [Stathis98], while the current through each trap is orders of magnitude higher because of the reduced tunnel distance. As a result, the number of traps participating in the SILC process becomes so small that one can easily distinguish the contribution of *each individual leakage path* [Degraeve05]. Both effects are depicted in Fig. 107.

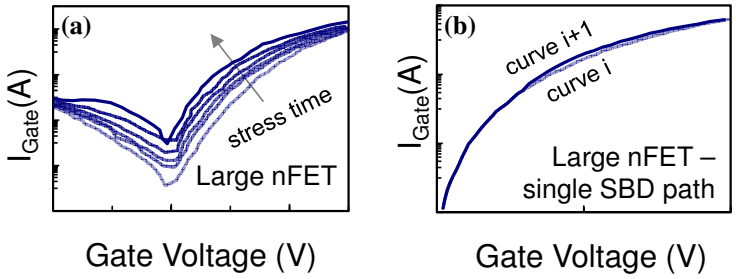



**Fig. 107: (a) On large devices, gate leakage current increases quasi continiously after subsequent gate oxide stress cycles. (b) It was shown by [Degraeve05] that on *ultra-thin oxide* devices, *each separate breakdown path* can have a *distinguishable impact* on the gate leakage current, shown as curve i before generation of a breakdown path and curve i+1 thereafter. [replotted from Cartier09 and Degraeve05].**

Fig. 108 shows the direct relation between SILC and the oxide trap density [Crupi04]. From these results it was proposed that SILC is caused by a trap-assisted conduction mechanism that involves one trap in the $HfO_2$. The deviation from 1/1 correlation close to breakdown, was attributed to two-trap

conduction paths. Due to a lack of correlation between the observed $V_{TH}$ shift and the SILC current, it was concluded at the time by Crupi *et al*. that the traps participating in the SILC mechanism are different from the ones that cause the hysteresis in the $I_D$-$V_G$ characteristics.

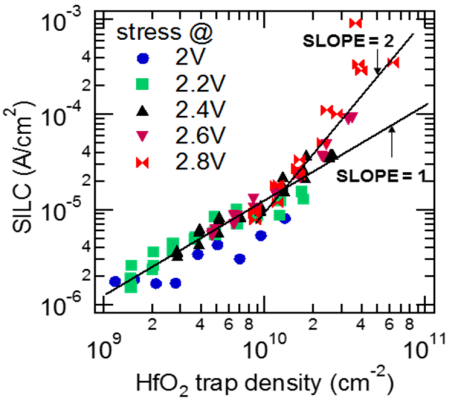



**Fig. 108: A 1/1 correlation for SILC and $HfO_2$ trap density (from charge pumping) at low trap density [replotted from Crupi04].**

## *4.3.2 The physical mechanism*

Stress induced gate leakage current is usually explained with trap-assisted tunneling, illustrated in Fig. 109. In this theory a trap in the oxide acts as a stepping stone for tunneling towards the gate electrode [Degraeve01]. A percolation path will be formed if there is an alignment of traps that connect cathode with the anode interface, and if the trap-to-trap and trap-to-interface distances are smaller than the percolation distance $x_{percolation}$. In other words, stress-induced current *steps* will be observed, each time a percolation path with is formed in the oxide. The magnitude of the current step depends on the percolation distance of this path: paths with a large $x_{percolation}$ will be poorly conducting and will result in a small current increase, paths, whereas paths with a small $x_{percolation}$ will be highly conducting. It can be easily understood that a higher oxide defect density will give rise to an elevated gate leakage current. It was shown by Degraeve that at low trap density, most of the current flows through single-trap paths. As the density of traps increases with

increasing stress time, the contribution of multi-trap conduction paths increases.

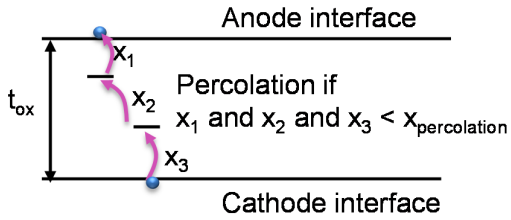



**Fig. 109: Schematic drawing of a gate current percolation path with two oxide traps [replotted from Degraeve01].**

The observed current steps are however not always accumulating. Sometimes elevated currents through the gate are also *appearing and disappearing continuously*. In that case, the current through the gate cathode $I_G$ will show an RTN-like behavior, with very similar time constants as what has been seen in $I_D$-RTN before.

### *4.3.3 $I_G$-RTN and $I_D$-RTN correlations*

Recent studies have focused on finding correlations between the gate current RTN and the drain current RTN [Toledano12, Chen11]. One of the conclusions was that in some cases, correlated $I_D$ and $I_G$ fluctuations could be observed (Fig. 110). This information could only be obtained on nanoscale devices which have also a percolated drain current.

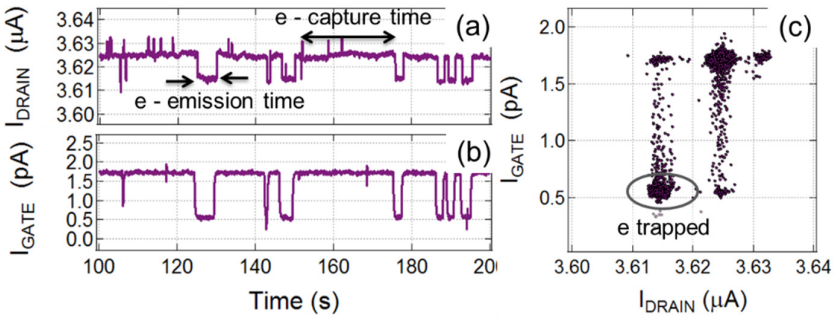



**Fig. 110: Observations of correlated (a) drain and (b) gate current fluctuations. (c) Two clusters of $I_D$/$I_G$ values are clearly visible. [replotted from Toledano12]**

A possible explanation for the observed correlation between $I_D$ and $I_G$-RTN was given by Toledano *et al.*: trapped BTI charges affect the band bending of the oxide, which on its turn affects the tunneling barrier for charges towards the gate electrode [Toledano12].

The crucial issue remaining is the mechanism through which traps can affect $I_G$. The large magnitude of $I_G$ fluctuations (about 75% of decrease in $I_G$) in Fig. 110 is striking. Although a charge capture does result in the local increase of the gate oxide tunneling barrier due to electrostatic screening [Franco12], the extent of these fluctuations suggest that this is a rather unlikely scenario. Indeed, it has been shown by Baumgartner *et al.* that electrostatic screening alone cannot alter the gate leakage current more than a few percent [Baumgartner13].

The same argument also rules out local electrostatic repelling of supply carriers in the injection electrode in the vicinity of the trapped charge as the origin of the fluctuations [Goes13].

Based upon the previous observations, it is clear that both BTI, RTN and SILC must be somehow correlated. Moreover, the correlated $I_G$ and $I_D$ current fluctuations in nanoscale devices indicate that BTI, RTN and SILC are multiple facets of the same source: defects in the gate dielectric (Fig. 111).

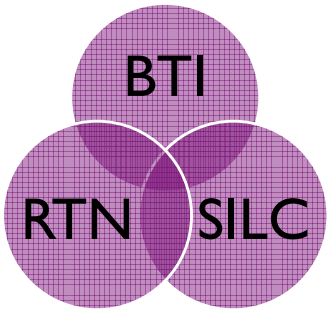



**Fig. 111: Illustration of how BTI, RTN and SILC degradation mechanisms are multiple facets of the same source: defects in the gate dielectric.**

### *4.3.4 Early models explaining BTI, RTN and SILC*

A theory *to explain the coupling* between the SILC, which was shown above to be based on trap-assisted tunneling (TAT) and BTI was proposed by Andersson *et al.* and later modified by Kaczer *et al.* and Goes *et al.* [Andersson90, Kaczer12, Goes13]. In this proposed theory, a trap in the oxide acts as a (thermally-activated) stepping stone for tunneling towards the gate electrode. This can be illustrated with a state-diagram (Fig. 112): carriers tunnel from injecting electrode 1s into metastable state 2' and either continue to the opposite electrode 1g, contributing to $\Delta I_G$, or a single carrier becomes captured in state 2, thereby disabling the conduction path and has its impact as fixed oxide charge on $I_D$.

In other words, the process involves *enhanced conduction* through the trap *when it is unoccupied by an electron*. Gate and drain current correlations can then be qualitatively constructed using this assumption, combined with the impact of a fixed charged trap.

In Section 4.4, we will run an extensive series of experiments to verify if this model is indeed able to capture all the observed correlations in $I_D$ and $I_G$ fluctuations and verify if this model is in agreement with the broadly accepted NMP model.

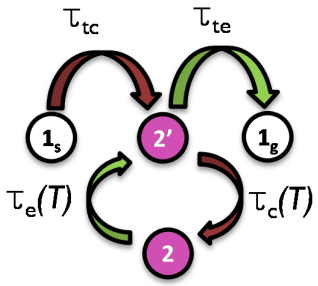



**Fig. 112: A state diagram describing the fluctuating $\Delta I_G$. Carriers tunnel from injecting electrode $1_s$ into metastable state 2' and either continue to the opposite electrode $1_g$, contributing to $\Delta I_G$, or a single carrier becomes captured in state 2, thereby disabling the conduction path [replotted from Kaczer13].**

## 4.4 Phenomenological study of SILC in nanoscale devices

This Section is structured as follows: first, we will describe our device selection and the experimental setup. Subsequently, we will analyze leakage paths, which are pre-existing in the device and develop a method to extract their position. Then we will define an experiment to find correlations between SILC and BTI. The first experiment reveals the properties of trapped-charge induced BTI-shifts, the gate voltage dependence of a single leakage current path on the total leakage current, and the effect of stress on the generation/activation of these gate leakage paths. In the next section, we will define a similar experiment, which reveals the time-dependent characteristics and correlations of these current fluctuations. Finally, we propose a *refined* 4-state NMP model that is able to capture all the observed effects.

### *4.4.1 Device selection and experimental setup*

For the measurements in this work we selected nanoscale nFETs with SiON and HKMG gate stacks respectively. The stacks were specifically selected to have a *comparable total physical thickness*, i.e. ~2.3nm. In this range, one

active defect is enough to enable a percolated current through the oxide, as described by [Degraeve05]. While the lateral dimensions of these devices are chosen small enough to increase the impact of single-defects on the drain current [Franco12], the gate leakage current density has to be sufficiently high to be within measurement resolution at gate biases around the device's threshold voltage.

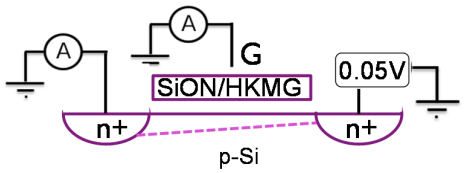



**Fig.** 113**: Illustration of the bias condition in a single nanoscale device used in this experiment. All the currents are monitors simultaneously, but the focus of this experiment goes out to drain and gate currents.**

All the FET currents were simultaneously measured with a pair of Keithley 2636 units, either with a voltage sweep or with a constant voltage at a rate of 10 samples/s. The measurements are performed at 25°C unless noted otherwise. The bias conditions are illustrated in Fig. 113.

### *4.4.2 Assessing leakage paths in pristine devices*

It is known that defects are pre-existing in the oxide. This means that even in pristine devices, TAT-leakage paths can already be apparent. The leakage paths are superimposed on direct tunneling gate leakage, as depicted in Fig. 114. By finding a device with no active TAT-paths, or, by actively enabling and disabling the TAT-path, its proper contribution to the leakage current, $\Delta V_{IG\_TAT}$ can be extracted.

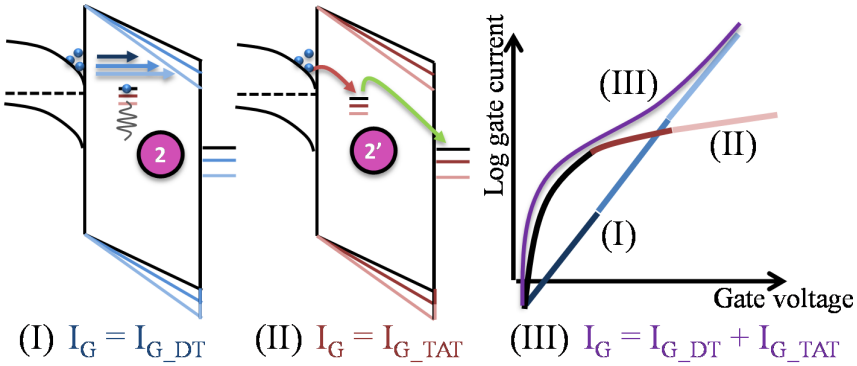



**Fig. 114: The schematic representation of the components of the gate leakage current. (I) A pre-existing or generated TAT-defect is charged and subsequently relaxed through multi-phonon emission. The trap cannot contribute to the gate leakage current and therefore only a direct tunneling current is measured. (II) The defect can actively contribute by trap-assisted tunneling (TAT). (III) This gate leakage path is superimposed on the direct tunneling gate leakage.**

The $I_G$-$V_G$ characteristics for both SiON as HKMG device (Fig. 115), show that many devices do not show a perfect exponential tunneling current. It is, however, possible to find devices with a 'defect-free' $I_G$-$V_G$, which shows a perfect exponential behavior within the measurement range (black curves in Fig. 115). This can indeed be expected based on the statistical model, described in Chapter 2, which predicts a Poisson spread of the defect occurrence. Since we expect only a few defects to occur in extremely scaled lateral dimensions, there is a reasonable probability of finding devices without any active defect. As a rough estimate, we assume 3.6 active defects per device $N_T$ (taken from experimental data from Weckx14). The probability of finding a device with *no active defects* is then 2.7%. It should be noted that this is under the assumption that apart from their defect configuration, the devices are identical.

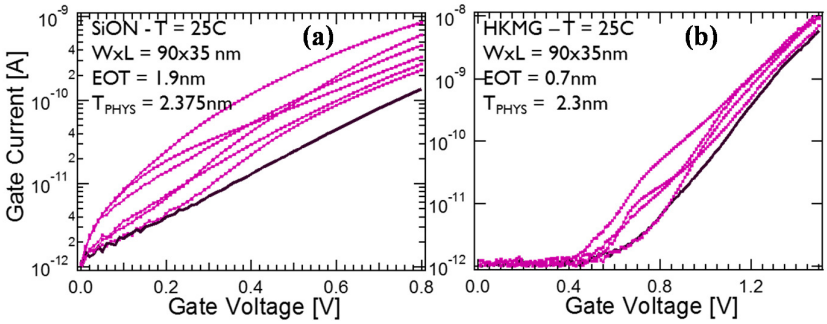



**Fig. 115: Initial $I_D$-$V_G$'s for different (a) SiON and (b) HKMG devices with similar physical thickness. In both cases, only a few devices show a near-perfect exponential leakage current (black curves). Other devices (purple curves) show a inflected $I_D$-$V_G$, indicating the presence of at least one TAT-leakage path. Source and drain are grounded in this case.**

An inflected gate leakage current (thus with a TAT-component) can be observed in both SiON as HKMG devices (Fig. 115), but remarkably, *it's not observed substantially more in the latter*, even though it is known that the high-k and $SiO_2$/high-k interface defect density is at least one order of magnitude higher. This is an indication that the enabling trap for the TAT is located in the $SiO_2$ rather than in the high-k material, a conclusion similar to [Bersuker11].

### *4.4.3 Position determination of the tunneling path*

When a small drain bias is applied (still in the linear regime) to the device, a voltage shift of the $I_G$ curves is observed (Fig. 116). The shift reflects the *lateral position of the TAT path* because the linear drop of the channel potential influences the local field in the gate oxide, and thus the leakage current through the TAT-path. Therefore, the ratio of voltage shift of this *single TAT-path* ($\Delta V_{IG\_TAT}$), with the applied $V_D$ thus determines the relative position of the trap in the channel $x_{trap}$ (Fig. 117):

$$x_{trap} = L_{channel} \frac{\Delta V_{IG\_TAT}}{V_D} \tag{4.1}$$

with $L_{channel}$ the length of the channel.

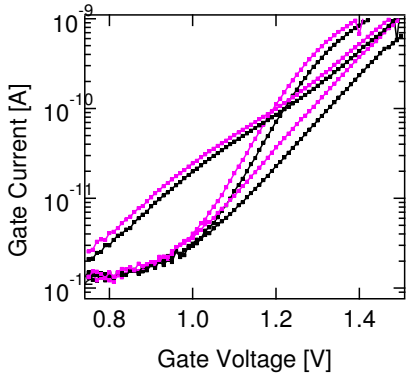



**Fig. 116: The $I_G$-$V_G$ with no drain bias (black) signal shifts if a (small) drain bias is applied (purple). This is due to the linear channel potential in this regime.**

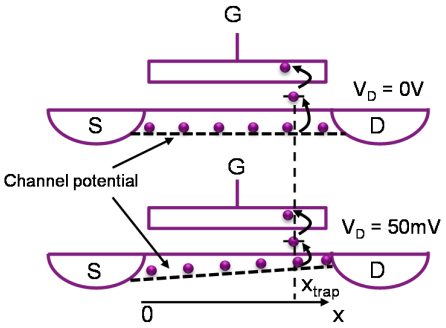



**Fig. 117: The linear drop over the channel potential allows to extract the lateral position of the TAT-paths $x_{trap}$, using the $V_D$ dependence of the $I_{G\_TAT}$ curves.**

This technique is a suitable alternative for the 's-ratio' technique in inversion [Crupi02], which relies on the *channel resistance* differences between the leakage path and the source and drain the junctions, which become unmeasurable in nanoscale and thus ultra-short FETs in inversion. The results utilizing the drain dependence of the TAT-paths for SiON and

HKMG devices, are depicted in Fig. 118. Interestingly, the SiON devices show clearly a more centered distribution of the TAT-paths than their HKMG counterparts. If this is not a processing related issue, this could be an indication that the extracted TAT-paths for the SiON FETs *require more than one oxide defect* a nearby horizontal location *to enable current percolation* through the gate oxide. The overall distribution would then no longer be uniform but normal distributed in the center. In HKMG devices, one oxide defect might still be sufficient for current percolation because of the lower tunneling barrier due to the lower bandgap of the high-k (Fig. 118), therefore revealing the uniform distribution of the oxide defects.

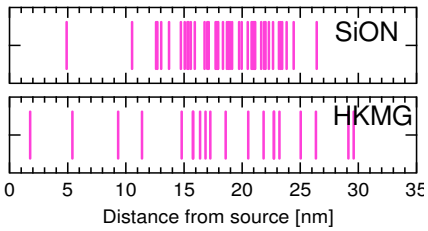



**Fig. 118: The SiON FETs (46 devices) show a more centered distribution of TAT-paths than the HKMG FETs (21 devices).**

## 4.5 The link between nanoscale SILC and BTI

In this Section, we will develop an experiment to reveal the properties of trapped-charge induced BTI-shifts in 4.5.1. Subsequently we study the gate voltage dependence of the current through a single leakage path (in 4.5.2), the effect of stress on the generation/activation of these gate leakage paths (in 4.5.3), the gate bias dependence of TAT activation (in 4.5.4) and their temperature dependence (in 4.5.5).

### *4.5.1 Experimental setup*

In a first step of the experiment, the devices are repeatedly stressed and relaxed for ~1s, and $I_D$-$V_G$ and $I_D$-$V_D$ traces are measured. After every 100 cycles, the stress voltage is increased by 100mV. The sequence of the experiment is depicted in Fig. 119. In this case, we select the HKMG devices

as those are known to be prone to PBTI, in contrast to SiON devices. The device size is 90x28nm (WxL), and the EOT = 0.9nm.

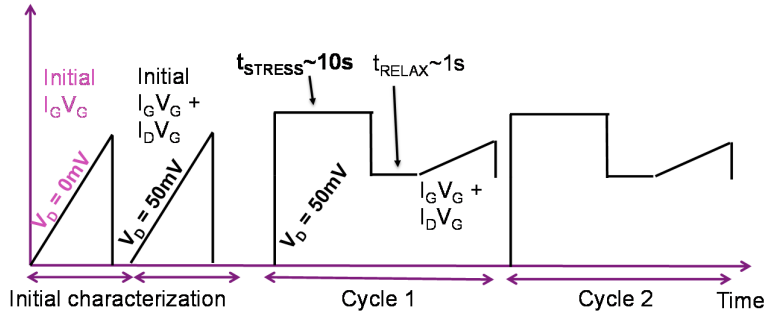



**Fig. 119: Cycles of BTI stress, relaxation, $I_DV_G$ and $I_DV_D$ sweeps are applied to the device. The total stress time is thus accumulated per cycle.**

### 4.5.2 *Partial correlation between $\Delta V_{TH}$ and $\Delta V_G$*

The result of the experiment is depicted in Fig. 120. The PBTI is visible as a degradation of the current after the subsequent stress cycles in Fig. 120 (a). Relaxation traces of the device after subsequent stresses shows that easily tens of charges must have been trapped, given the small step size of each trapped charge in the $I_D$. Note that we will systematically utilize a the darker shade of purple for the $I_D$-$V_G$ or $I_G$-$V_G$ trace to indicate that the device has endured more accumulated stress.

This is also reflected as a shift of the $I_D$-$V_G$ characteristic (a positive $\Delta V_{TH}$) in Fig. 120 (b). If we look at the $I_G$ caused by direct tunneling (denoted as $I_{G\_DT}$) in the same picture, these currents are also shifted towards a more positive $V_G$ after the subsequent stress cycles. From the inset in Fig. 120, it can be observed that both the $I_D$ as the $I_{G\_DT}$ a comparable total shift. We also observe that in this particular case, no TAT-paths have been created. We conclude from this, that the $I_{G\_DT}$ are also shifted by the electrostatic charge accumulated in the oxide due to the BTI-induced charge trapping. The inset however also reveals that *there is no 1/1 correlation* between the shifts of the gate leakage currents ($\Delta V_G$) and the drain current shift $\Delta V_{TH}$.

The lack of the 1/1 correlation between the impact of the defects on the drain current (i.e. the position w.r.t. the percolation path) and the screening of the gate current can be explained: the trapped BTI-charges will have an electrostatic impact on the TAT-path, but not the same as their impact on the $I_D$. In other words: a trap close to a channel percolation path (causing a large $\Delta I_D$) will not necessarily induce a large voltage shift on the leakage current and vice versa (Fig. 121). Because of the large number of charges that were trapped during this experiment, the *total* $\Delta V_{TH}$ of ~70mV corresponds however roughly to the $\Delta V_{G\_DT}$.

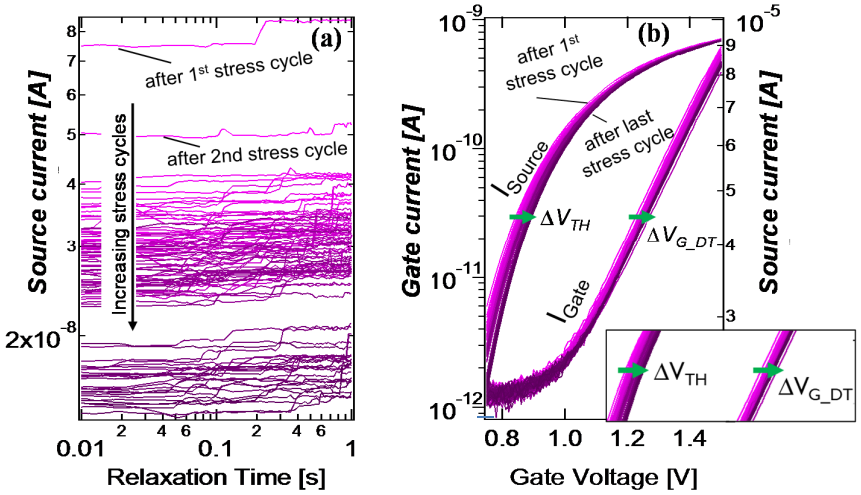



**Fig. 120 (a) The proposed experiment performed on a HKMG device (WxL: 90x28nm, EOT = 0.9nm) for one stress condition, repeated 100 times. (b) The $I_D$-$V_G$ and $I_G$-$V_G$ traces after the short relaxation.**

In other devices, TAT-paths (appearing as sudden bumps in the sudden steps in the $I_G$-$V_G$ characteristic) will be generated during the stress. In those cases, the appearance of the TAT-trace with convoluted with the electrostatic shift caused by the oxide charges. In order to de-convolute the TAT-paths, we compensate the $I_G$-$V_G$ for the shift measured in $I_D$. This can be done individually for every stress cycle, as the $I_D$ and $I_G$ are measured simultaneously.

To illustrate this principle, we depict in Fig. 122 the measured BTI-induced $\Delta V_{TH}$ extracted from the $I_D$-$V_G$ traces (in this case for another device in the same experiment). Also here, $\Delta V_{TH}$ increases rather gradually per stress cycle, because of the high defect density in the high-k, which reduces the *average impact per trap* $\eta$, as discussed in Chapter 2. A $\Delta V_{TH}$ of 40mV after the first stress cycle $V_{STRESS}$ = 1.7V is observed, and about 70mV at the end of this set of stress cycles. After the first stress cycle at 1.8V, the $\Delta V_{TH}$ has already reaches 75mV.

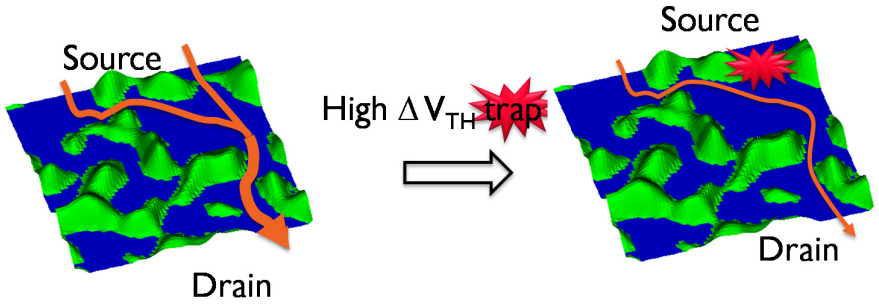



**Fig. 121: The position of the trap in the horizontal plane determines the impact on the $V_{TH}$, whereas the position in the vertical plane will determine its gate leakage current.**

In a next step, the obtained $\Delta V_{TH}$ from Fig. 122 are then utilized in Fig. 123 to superimpose this BTI-induced shift on the $I_G$ traces. As a result, the additional TAT-paths become visible in Fig. 123 (b), proving that the correlation between the $\Delta V_{TH}$ and the $\Delta V_G$ should be taken into account. The result, however, also shows an increased dispersion in the $I_G$ traces, because of the lack of 1/1 correlation between the $\Delta V_{TH}$, used to correct, and the $\Delta V_G$.

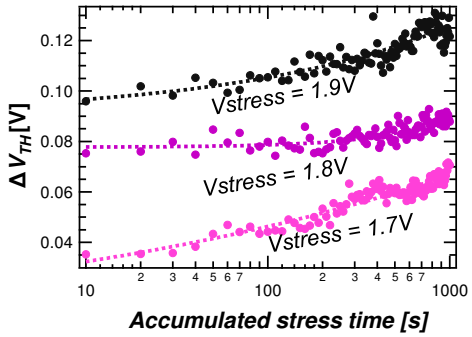



**Fig. 122: Extracted BTI-induced $\Delta V_{TH}$ at 0.01s of relaxation plotted against the total accumulated stress time. The colors indicate 3 different stress voltages that were applied *on the same device*. Note that we reset the cumulative stress time after each stress voltage.**

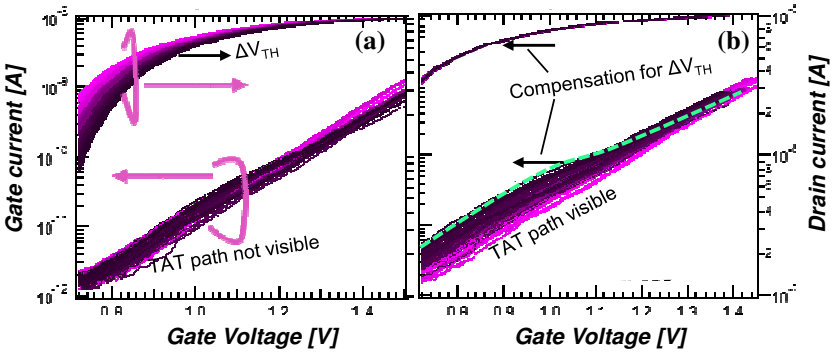



**Fig. 123. The $V_{TH}$ shifts is quasi-continuously with increasing stress cycles and voltages (darker traces means more accumulated stress), visible in the $I_D$, due to the large number of trapped charges. The $I_G$ shows mostly discontinuous steps which are not distinguishable in $I_D$. In (a) the trapped-charge screening effect on the TAT is apparent, while in (b) correcting for this $V_{TH}$ shift, also illustrated in $I_D$, makes the TAT paths visible.**

### 4.5.3 Ambiguity of stress on SILC current

The above described experiment was repeated on multiple devices. A particular case is depicted in Fig. 124, which shows that the TAT-paths can not only be generated or activated by stress as in (a), but *also de-activated* for longer times, as in (b). A similar effect on drain-like RTN was reported earlier by Grasser et al., i.e. that some *BTI and RTN defects "... completely disappear and reappear over extended time intervals*" [Grasser13]. The ambiguous effect of the stress bias on the SILC or TAT-current was not yet described in literature.

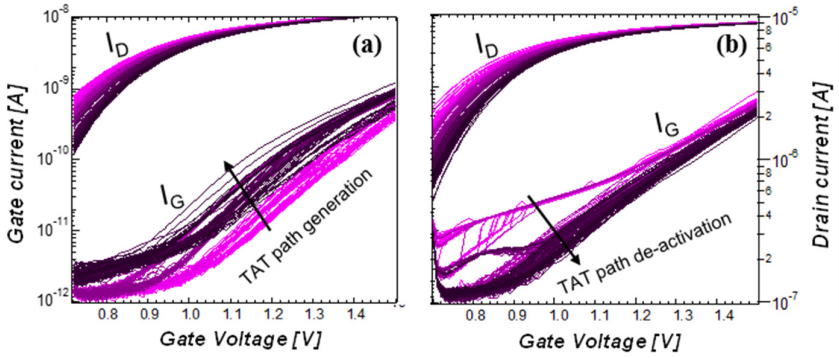



**Fig. 124. $I_G$ shows mostly *discontinuous* steps which are not distinguishable in $I_D$. In (a) TAT paths are *activated* after stress whereas in (b) the SILC paths are *de-activated* after the stress phases.**

### 4.5.4 Gate bias dependence of TAT-path activation

Another particular device is depicted in Fig. 125. In this case, the TAT-paths are activated and de-activated *dynamically during* the $I_G$-$V_G$ sweeps. It is well known that the *generation of defects* (i.e. thus potential new TAT-paths) is strongly gate bias dependent because a high stress voltage will induce new defects [Degraeve01]. In this case however, the activation and de-activation of *one* particular path shows a clear gate bias dependence during the sweep, as extracted in Fig. 145 (b). In this case, the probability for disabling the strongest of both TAT-paths increases with gate bias.

Remarkably, for this device, the corresponding $I_D$-$V_G$ trace is not showing any RTN-like modulation within the measurement resolution. To investigate correlated RTN, we will focus on constant voltage measurements in Section 4.6.

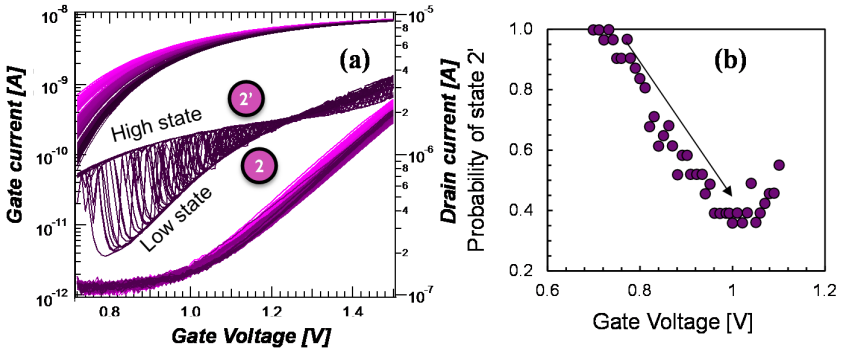



**Fig. 125: (a) Generation and oscillation between two TAT states and (b) the gate voltage dependence of occupation probability of the each of the TAT states.**

## *4.5.5 Temperature dependence of TAT-path activation*

The above described experiment was subsequently repeated at elevated temperature. In this case, 125°C, which is the typical measurement condition for BTI stress. At this temperature, the $I_G$ traces become unstable and vary strongly from sweep to sweep (Fig. 126). The instability in $I_G$ is, however, not visible in $I_D$. More charge trapping—observed as $V_{TH}$ shift in $I_D$—is apparent. Shifts up to 300mV are observed (not shown here), although the device remains fully functional.

Meanwhile, the gate leakage current is strongly increased to the point where the individual TAT-paths become indistinguishable. Their activation and de-activation time constants are strongly temperature dependent and decrease below measurement integration time. This temperature activation is consistent with the phonon relaxation model, illustrated earlier in Fig. 112, which states that the transition of state 2 (an active leakage path) to state 2' (an inactive leakage path) is temperature activated [Kaczer12]. Indeed, we observe here

that the activation and de-activation of the leakage paths occur more promptly at elevated temperatures.

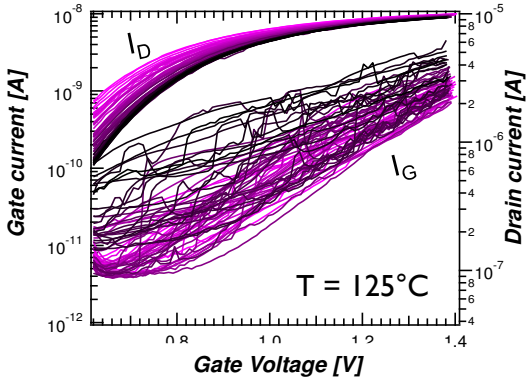



**Fig. 126: At high temperatures, the 2'↔2 transitions are occurring promptly and faster than the measurement integration time.**

From all of the above described experiments, no 1/1 correlations between individual $I_G$ and $I_D$ shifts were found. For example, the activated TAT-paths resulting in a large $\Delta I_G$ did not result in a large $V_{TH}$ shift. Correlations in smaller shifts in $I_G$ and $I_D$ are harder to observe because of measurement noise when performing a sweep of $V_G$. Therefore, in order to observe *directly correlated* instabilities, we will focus towards *constant bias* measurements in Section 0.

### *4.5.6 Conclusions on experimental data*

Based upon the dataset of our earlier proposed measurement, we can conclude that *on average* there is a correlation between the $\Delta I_G$ and $\Delta I_D$, because they are both impacted by the electrostatics of trapped charges, but a lack of direct correlation because of the percolated nature of the drain current. We showed that TAT-paths can not only be activated but also de-activated by stress, i.e. the defects responsible for TAT can be present but in a de-activated state. We showed that gate-bias can alter the probability of the activating TAT-paths. Finally, it was shown that the activation and de-activation of the TAT paths is strongly temperature activated, which is consistent with the

proposed model of Kaczer *et al* [Kaczer12]. Before we provide and discuss our refinement of the NMP physical model in Section 4.7, able to explain the above measurement results, we will discuss a second set of measurement results which focuses on finding correlated $I_G$ and $I_D$ RTN in Section 4.6.

## 4.6 Gate and drain RTN correlations

Focusing on the time evolution of gate and drain currents might help to find correlations between $\Delta I_G$ and $\Delta I_D$. Some research groups already noted a *positively* correlated $I_G$ and $I_D$ in *pFET* devices [Toledano12], whereas other groups have reported *both negatively and positively correlated RTN* in nFET devices [Ji13, Liu14].

### *4.6.1 Experimental setup*

By measuring $I_D$ and $I_G$ simultaneously for an extended time correlations in $I_G$ and $I_D$ can easily be captured. We add a short stress phase in between to stimulate activation or de-activation of certain TAT-paths. In this case, we opted for the SiON transistors to minimize the number of drain steps with respect to the number of active TAT-paths. Note that typically the defect density of SiON devices is 10x less than in HKMG [Degraeve05].

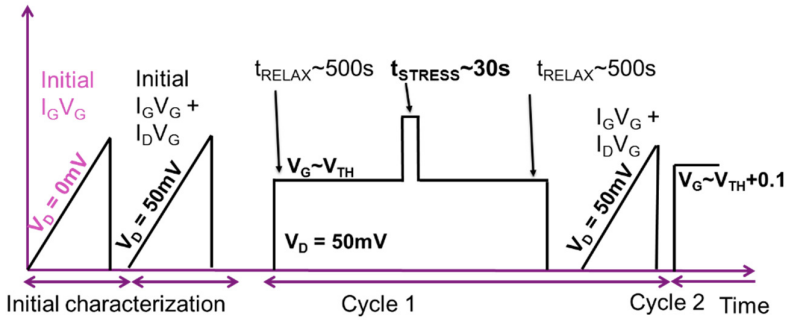



**Fig. 127: This experiment focusses on the time evolution and correlation of both the gate and drain current after one short stress phase. During and extended $t_{RELAX}$, correlations between RTN in $I_G$ and $I_D$ are measured, as well as $I_B$.**

### 4.6.2 Observations of directly correlated $\Delta I_G$ and $\Delta I_D$

Analyzing the result of the experiment, we observed that most of the RTN events in the drain show no effect on the gate current and vice versa. This is remarkable as it is generally assumed that PBTI is primarily caused by traps in the high-k and the $SiO_2$/high-k interface [Weckx13], which is also a more favorable position than the $Si/SiO_2$ interface (where NBTI traps are typically located) from the perspective of trap-assisted tunneling.

Apart from the discharging events after the stress pulse due to non-equilibrium of the system, as described by [Grasser13], single and multi-level RTN signals are visible, mostly in $I_D$. These signals are ideal for studying correlations in TAT-paths and BTI traps. The multi-level RTN traps can be distinguished and treated as separate traps, as long as the step height differences are large enough to discriminate. In a few cases, correlated gate and drain RTN is observed, and in a few cases both *positively and negatively correlated RTN could be observed* within the same device. An example these traces are depicted in Fig. 128.

The RTN on the $I_G$ (thus caused by switching TAT-paths) can now be observed when the gate leakage current is measured at constant bias. This proves once more that *one* trap can be responsible for activating or de-activating these leakage paths. As the single path $\Delta I_G$ can be $\approx 2.5 \times 10^{-11}$ A (also derived from Fig. 128) about $1.5 \times 10^{8}$ charges/s should be captured and emitted to the gate by the oxide trap. The slowest time constant in the inelastic TAT process (thus either $\tau_{tc}$ and $\tau_{te}$, as will be discussed later) is thus is no larger than $\sim 1 \times 10^{-9}$ s.

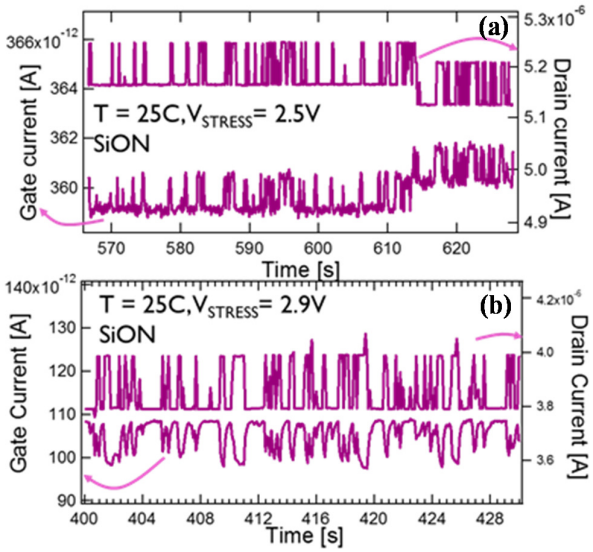



**Fig. 128. Both (a) positive and (b) negative $I_D/I_G$ correlations are measured on separate SiON nFETs after stress. The negative correlation in (b) also shows signs of a second RTN path in the drain current, and is superimposed both on the low and the high state of the signal, but not visible in the gate current.**

Even though it was already stated that the electrostatic model cannot account for the large gate leakage fluctuations seen in earlier experiments, the observation of negatively correlated RTN is direct evidence that disagrees with the intrinsic polarity requirement of the electrostatic model.

It should however be noted that the relative impacts of the observed correlated events are rather low. This is agreement with earlier experiments from Toledano-Luque et al. (Fig. 129).

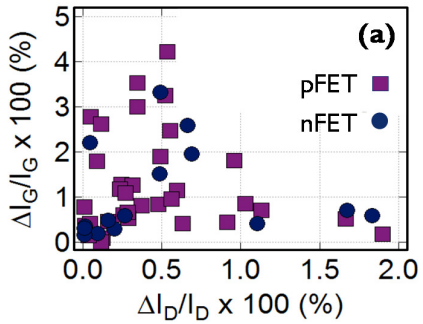



**Fig. 129: Correlation plot of relative changes in $I_D$ and $I_G$. Note that upper right corner (large $I_D$ and large $I_G$ fluctuations) is unpopulated. The mechanism that was proposed involves enhanced conduction through a gate oxide trap when it is unoccupied. Such a process would be most efficient for traps close to the center of the oxide, readily explaining why there are no defects that cause *both* large $\Delta I_G$ and large $\Delta I_D$ [replotted from Toledano12].**

The lack of defects causing *both* large $\Delta I_G$ and large $\Delta I_D$ can be readily explained as follows: the impact of a charged trap on $I_D$ depends on the distance of the trap from the critical point of a source-drain percolation path as discussed earlier. A TAT process would however be most efficient for traps close to the center of the oxide [Ielmini02].

## 4.7 Unified defect-based model for BTI, RTN and SILC

In this Section, we will describe a unified defect-based model that explains the current correlations in BTI, RTN and SILC.

Our model is schematically described in Fig. 130. It can describe *no*, *positive* and *negative* $I_G/I_D$ correlations for nFET devices, and is fully consistent with the 4-state defect model as proposed by [Grasser12]. The gist of the model is the *net observable charge*, which is defined by the *product of the occupancy probability and the defect charge state*.

Fig. 130 shows how the occurrence of certain microscopically observable events can be depicted following the convention of the 4-state defect model:

- BTI charge trapping from a not-yet trapped charge in the channel (state $1_s$) over a metastable state (state 2') into a stable and reconfigured state (state 2).
- Trap assisted tunneling from a charge in the channel (state 1s) over an active trap (metastable state 2') towards the gate electrode (state 1g).
- *Inhibited* trap-assisted tunneling over the reconfigured and charged defect, because the charge in the trap is fixed in a stable state 2.

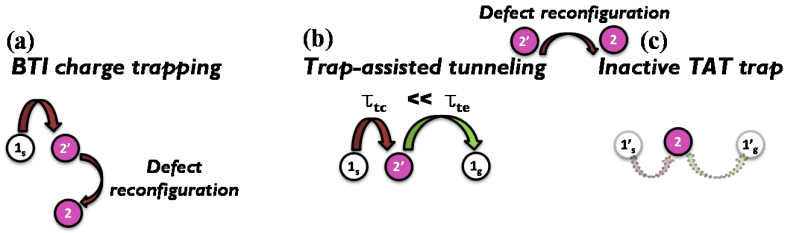



**Fig. 130: Various microscopically observed phenomena explained with the stable and metastable states of the NMP model. (a) BTI charge trapping over a metastable state into a stable and reconfigured state. (b) Trap assisted tunneling over an active trap and (c) inhibited trap-assisted tunneling over the reconfigured and charged defect. Note that the blank circles indicate a net neutral charge of the trap towards the channel, whereas a purple filled circle indicates a net positive or negative charge.**

### *4.7.1 Absence of correlations of $\Delta I_D$ and $\Delta I_G$*

The fluctuations in $I_G$ which are not linked by a fluctuation in $I_D$, can be explained with an active TAT-path located closer to the channel (Fig. 131). As discussed above, the net charge is defined by the product of the occupancy probability and the defect charge state. Therefore, even in its metastable position, this trap will have a net charge, as in this case, the time constant for inelastic tunneling from the channel towards the trap $\tau_{tc}$ (1s→2') is smaller than the time constant to tunnel from the trap towards the gate $\tau_{te}$ (2'→1g), because of its position. Moreover, the trap's reconfigured stable state 2 is by definition

also a charged defect state and inhibits carriers to hop towards the gate electrode. Therefore, *whilst activating or de-activating the TAT-path*, *no net charging in the oxide occurs.* As a result, no impact on the drain current is expected, regardless if the TAT-path is activated or not.

The time constants for reconfiguration from the defect by NMP (2'→2) is defined as $\tau_c$, whereas the time constant for phonon absorption towards its stable state (2→2') is defined as $\tau_e$. Finally, it is assumed that in this case, the trap-assisted transitions from the channel towards gate electrode *over the reconfigured defect* are strongly unfavorable, i.e. either $\tau'_{tc}$ or $\tau'_{te}$ are very large.

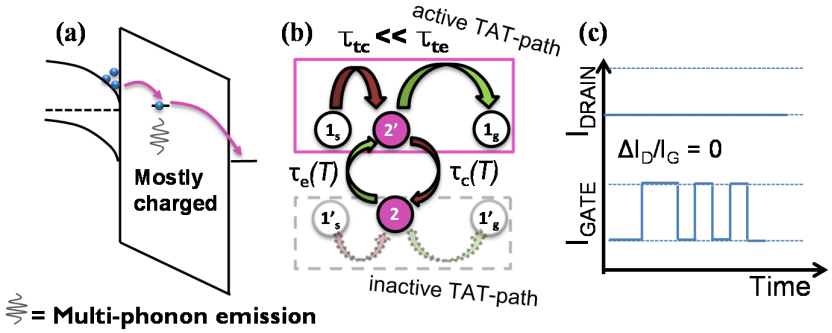



**Fig. 131. Model for TAT explaining the *absence* of $I_D/I_G$ correlations for nFET devices. (a) Band diagram of TAT tunneling and relaxation of the defect after charge trapping. (b) State diagram of carriers hopping over the trap and possible relaxation after charge trapping but activation or de-activation of the TAT-path has no impact on the net charge of the defect. (c) Illustration of the resulting $I_D$ and $I_G$.**

## *4.7.2 Positively correlated $\Delta I_D$ and $\Delta I_G$*

The obtained positively correlated RTN, which were already experimentally shown by [Toledano12] and [Chen11] and are also observed in our measurements, can be explained by cases where the *capture time is dominant* (Fig. 132). This mean that the transition 1s→2' is slower than the transition between 2' and 1g, i.e. that $\tau_{tc} << \tau_{te}$. A physical explanation for

this could be that the trap is located far from the interface. The trap is then quasi empty, as every channel carrier that was able to arrive, will immediately propagate towards the gate electrode.

In contrast to 4.7.1, in this case, *phonon relaxation* occurring in the TAT-defect (transition from 2'→2) *will result in a net charging effect* of the gate oxide. The net charge will on its turn be observable in $I_D$, because the net charge when the TAT-path is active is 0, whereas it becomes fixedly charged whenever the defect changes into its stable state 2. The net charge will reduce the current of the $I_D$, and will also avoid that the TAT-path becomes active, thus also reduce $I_G$. This explains how a positive RTN correlation can be observed ($\Delta I_D/\Delta I_G > 0$).

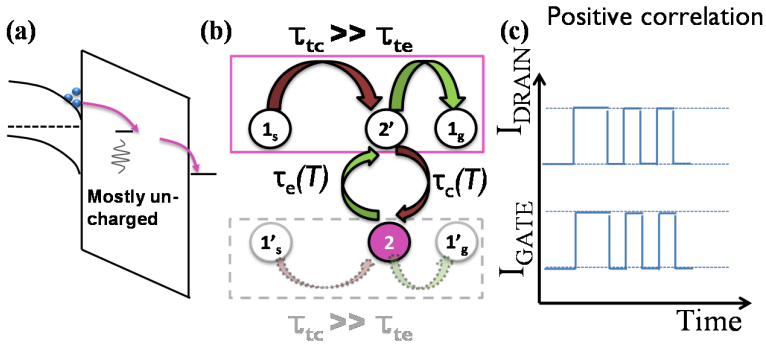



**Fig. 132. Model for TAT explaining the positive $I_D/I_G$ correlations for nFET devices. If $\tau_{tc} >> \tau_{te}$ (for states far away from the interface), state 2' is mostly unoccupied. In this case, the phonon relaxation of the TAT defect towards state 2 *will result* in a net charge difference observable in $I_D$. We therefore find a *positive* RTN correlation ($\Delta I_D/\Delta I_G > 0$).**

## *4.7.3 Negatively correlated $\Delta I_D$ and $\Delta I_G$*

Also the *negative* correlation observed earlier in [Ji13, Liu14] and now seen here in pFET devices, cannot just be explained with a TAT scheme and neither with direct electrostatic interaction. A plausible explanation for these correlations is by favorable transitions *between the secondary defect states* $1'_g/1'_s$ and 2 (Fig. 133). Experimental observation of the capability of secondary defect states to conduct current was shown in Fig. 125. In this case,

this means that both transition rates $\tau'_{tc}$ or $\tau'_{te}$ *between the secondary states* should be much lower.

Conditions that have to be met in order to explain this above effect, are that:

- the transition rates *between the secondary states* are such that $\tau'_{tc} << \tau'_{te}$, to obtain the net charging effect when the TAT-current is enabled,
- the transition rates are *opposite for the primary states*, i.e. $\tau_{tc} >> \tau_{te}$, necessary to have a net discharged state when the trap is in state 2' and the TAT-current is disabled.

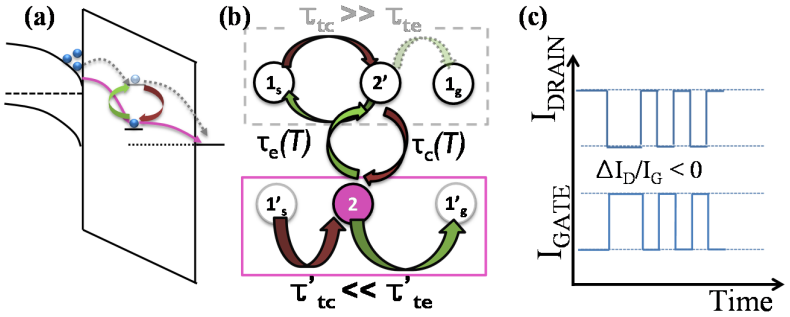



**Fig. 133. Model for TAT explaining the negative $I_D/I_G$ correlations for nFET devices. State 2 (the defect state after reconfiguration and after charge trapping) *also has to be a net charged state* to impact $I_D$. Thus, for inverse correlation, the transition rates $\tau'_{tc} << \tau'_{te}$ between the secondary states are 1) much shorter and 2) opposite in their magnitudes: $\tau_{tc} >> \tau_{te}$.**

One other condition that should be met in order for this state model to explain the above effects, is that state transition rate 2→2' ($\tau_c$) should be smaller than transition rate 2→1'g ($\tau'_{te}$). This is obviously the case as the time constants for TAT are extremely low (below nanoseconds, discussed earlier), whereas $\tau_c$ is a NMP-activated transition with switching events measureable during the sweep, even at elevated temperatures (Fig. 126).

Finally, the earlier observed gate voltage dependence of TAT-path activation, e.g. the one observed in Fig. 125 can be ascribed to a bias-dependent occupancy probability of state 2' influencing the 2'→2 transition.

In other words, even though the NMP relaxation is expected to be gate bias independent, the leverage of the typical gate bias dependency of BTI parameters $\tau_{tc}$ and $\tau'_{tc}$ will cause a change in the occupation probability of states 2 and 2' respectively (depending if it is a positively or negatively correlated path). And, as discussed before, the occupancy of state 2 is determined by its *effective capture time*, which is determined by the product of the occupancy of metastable state 2' and the time for passing from metastable state 2' towards stable state 2.

## 4.8 Conclusions

Extended analysis of currents on all terminals of nanoscaled devices yields significant insight in TAT/SILC, RTN and BTI mechanisms. We proposed a model capable of explaining the measurement observations, and in particular the positive and negative $I_D$ and $I_G$ correlations. It was found that the multi-state defect model can explain correlated and non-correlated gate currents.

Also, a new methodology was shown that allows the extraction of the physical position of the leakage paths *in inversion regime*, in contrast to the s-ratio technique which only works in accumulation for nanoscale devices.

Finally, it was found that in HKMG, most "BTI visible defects" do not have a large contribution to SILC current and vice versa, SILC defects do not show a large contribution to the $V_{TH}$ shift. Even though stress can result in defect generation, it will not necessarily result in activation of leakage paths, but it can also de-activate these paths.

## 4.9 References

[Andersson90] Andersson M.O., Xiao Z., Norrman S., and Engstrom O., "Model based on trap-assisted tunneling for two-level current fluctuations in submicrometer metal—silicon-dioxide—silicon diodes", Phys. Rev. B., Vol 41, pp. 9836-9842, (1990).

[Baumgartner13] Baumgartner O. et al., "Direct Tunneling and Gate Current Fluctuations", in Proc. SISPAD, pp. 17-20, (2013).

[Bersuker11] Bersuker G. et al, "Mechanism of high-k dielectric-induced breakdown of the interfacial SiO2 layer", in Proc. IRPS, pp. 373-378, (2011).

[Cartier09] Cartier E. and Kerber A., "Stress-Induced Leakage Current and Defect Generation in nFETs with HfO2/TiN Gate Stacks during Positive-Bias Temperature Stress", in Proc. IRPS, pp. 6.2.1-4, (2009).

[Chen11] Chen C-Y. et al., "Correlation of Id- and Ig-Random Telegraph Noise to Positive Bias Temperature Instability in Scaled High-κ/Metal Gate n-type MOSFETs", in Proc. IRPS, pp. 190 – 195, (2011).

[Crupi02] Crupi F. et al., "Location and Hardness of the Oxide Breakdown in Short Channel n- and p-MOSFETs", in Proc. IRPS, pp. 55-59, (2002).

[Crupi04] Crupi F. et al, "Correlation between Stress-Induced Leakage Current (SILC) and the HfO2 bulk trap density in a SiO2/HfO2 stack" in Proc. IRPS, pp. 181-187, (2004).

[DeBlauwe96] De Blauwe J. et al., "Study of DC Stress Induced Leakage Current (SILC) and its Dependence on Oxide Nitridation" in Proc. ESSDERC, pp. 361-364 (1996).

[Degraeve01] Degraeve R. et al., "Statistical model for Stress-Induced Leakage Current and pre-breakdown current jumps in ultra-thin oxide layers", in IEDM Tech. Dig., pp. 121-124, (2001)

[Degraeve05] Degraeve R. et al., "Measurement and statistical analysis of single trap current-voltage characteristics in ultrathin SiON", in Proc. IRPS, pp. 360-365, (2005).

[Franco12] Franco J., et al, "Impact of Single Charged Gate Oxide Defects on the Performance and Scaling of Nanoscaled FETs", in Proc. IRPS, pp. 5A4.1-4 , (2012).

[Goes13] Goes W. et al, "Understanding correlated drain and gate current fluctuations", in Proc. of IPFA, pp. 51-56, (2013).

[Grasser10a] Grasser T., et al, "The Paradigm Shift in Understanding the Bias Temperature Instability: From Reaction–Diffusion to Switching Oxide Traps," in IEEE Transactions on Electron Devices, vol. 58, iss. 11, pp. 1–15, (2011).

[Grasser10b] Grasser T., et al., "The time dependent defect spectroscopy (TDDS) for the characterization of the bias temperature instability", in Proc. IRPS, pp. 16-25, (2010).

[Grasser12] Grasser T. et al., "On the Microscopic Origin of the Frequency Dependence of Hole Trapping in pMOSFETs," in Proc. IEDM, pp. 19.6.1–19.6.4, (2012).

[Grasser14] Grasser T. et al., "Characterization and Modeling of Charge Trapping: From Single Defects to Devices", in 2014 IEEE International Conference on IC Design & Technology, pp1-4, (2014).

[Ielmini02] Ielmini D. et al., "A statistical model for SILC in flash memories" in IEEE Trans. Electron Devices, Vol. 49, pp. 1955–1961, (2002).

[Ji13] Ji X. et al., "The Physical Mechanisms of Ig Random Telegraph Noise in Deeply Scaled pMOSFETs," in Proc. IRPS, pp. XT.7.1–XT.7.5, (2013).

[Kaczer08] Kaczer B. et al, "Ubiquitous Relaxation in BTI stressing—New Evaluation and Insights", in Proc. IRPS, pp. 20-26, (2008).

[Kaczer09] Kaczer B. et al., "NBTI from the perspective of defect states with widely distributed time scales", in Proc. IRPS, pp. 55–60, (2009).

[Kaczer13] Kaczer B. et al., "Gate Current Random Telegraph Noise and Single Defect Conduction", in Microelectronic Engineering Volume 109, pp. 123–125, (2013).

[Kauerauf11] Kauerauf T. et al., "Methodologies for sub-1nm EOT TDDB evaluation" in Proc. IRPS, pp. 7-16, (2011).

[Kirton89] Kirton M. and Uren M, "Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency (1/*f*) noise", in Adv. Phys. 38, 367, (1989).

[Liu14] Liu W. et al, "Analysis of Correlated Gate and Drain Random Telegraph Noise in Post-Soft Breakdown TiN/HfLaO/SiOx nMOSFETs", in Electron Devices Letters, Vol. 35, No. 2, pp. 157-159, (2014).

[McWhorter57] McWhorter A.L., "1/f Noise and Germanium Surface Properties," in Sem. Surf. Phys, pp. 207–228, (1957).

[Sasse08] Sasse G.T. and Schmitz J., "Application and Evaluation of the RF Charge-Pumping Technique," in IEEE Transactions on Electron Devices, vol. 55, no. 3, pp. 881-889, (2008).

[Stathis98] Stathis J.H. and DiMaria D.J., "Reliability projection for ultrathin oxides at low voltage", IEDM Tech. Dig., pp, 167-170, (1998).

[Toledano12] Toledano-Luque M., et al., "Correlation of single trapping and detrapping effects in drain and gate currents of nanoscaled nFETs and pFETs", in Proc. IRPS, pp. XT. 5.1 – XT. 5.6, (2012).

[Weckx15] Weckx P. et al., "Characterization of time-dependent variability using 32k transistor arrays in advanced HK/MG technology", in Proc. IRPS, pp. 3B.1-6, (2015).

# Chapter 5: Assessing self-heating effects in scaled MOSFET nodes

In this Chapter, we develop and assess various methodologies to study self-heating effects in scaled MOSFET transistors, both experimentally as well as through modeling efforts. A case study will be presented assessing the self-heating effects in planar, FinFET and GAA-NW devices.

## 5.1 Introduction

When changing the device geometry from planar to multi-gate such as FinFETs, the concern of self-heating effects (SHE) grows. Also for SOI (Silicon On Isolator) devices, this self-heating effect is a concern, due to the significantly smaller thermal conductivity of silicon dioxide ($\kappa_{SiO2} = 1.40$ $WK^{-1}m^{-1}$) compared to that of bulk silicon ($\kappa_{Si} = 148$ $WK^{-1}m^{-1}$) at room temperature [Dallman95], [Fiegna08].

Dissipated power in the device leads to a temperature rise that can be non-uniform and cause local hot-spots. The high temperature of a device can significantly impact the device performance and reliability. Typically the *drive current decreases* with temperature, and the local temperature elevated by self-heating can additionally accelerate device degradation, which *impacts transistor reliability* and safety margins including BTI, trap assisted leakage, TDDB and CHC degradation.

As discussed in Chapter 2, the transistor self-heating effect can be measured or simulated in numerous ways. Measurements can be either direct, i.e. by directly sensing the $\Delta T$ of the device (e.g. gate resistance thermometry) or indirect, i.e. by analyzing the time and temperature dependence of the drive current (e.g. by RF-measurements) [Beppu13, Scholten09]. The simulation can range from employing classical Fourier's law of diffusion and solving linear heat equations, to using non-equilibrium statistical mechanics by modeling electron-phonon interactions and solving phonon energy balance equations or the phonon Boltzmann transport equation [Vasileska12, Rhyner13].

In this Chapter, we first try to utilize existing SHE measurement techniques on our devices, subsequently we introduce a new methodology to measure the

SHE in planar, FinFET and nanowire devices. Subsequently, we introduce various simulation methodologies to capture the effect. Thereafter, we make a case study of self-heating effects in planar transistors versus FinFET transistors, and latest-generation FinFETs versus GAA-NW transistors. We will also make predictions of the self-heating effects in future Si-based nodes. In the last part of this Chapter, we briefly discuss the effects of self-heating on circuit performance and reliability.

## 5.2 Applying existing methodologies to planar devices

In this Section, in order to assess the thermal resistance $R_{TH}$ and capacitance $C_{TH}$ of nanoscaled planar devices, we will apply the existing Pulsed-IV and high frequency conductance technique by RF-measurements.

### *5.2.1 Pulsed-IV methodology*

To assess the self-heating effect in our planar devices, we start by utilizing the PIV-technique which was introduced in Chapter 2. Fig. 134 shows the result of a PIV-measurement with 150ns pulse width. Various instances on a device show that PIV measurements exhibit either similar or lower drive currents than DC measurements. Also the charge trapping and de-trapping is leading to two distinct $I_D$-$V_G$'s for instance #2.

From our measurements on nanoscaled planar FETs, it appears thus that, even in strong saturation, the drive current during pulsed measurements is not only *similar* or *lower* than the DC measured current, but also strongly affected by charge trapping. It can of course be expected that charges are trapped during these strong saturation conditions, i.e. in the HC regime. It is thus clear that charge trapping effects should be de-convoluted from the SHE to obtain the actual temperature. From these measurements alone, we are thus unable to determine the impact of self-heating on the drive current, nor the temperature of the devices.

A similar experiment was performed on pFinFETs, which are expected to show enhanced self-heating, and is shown in Fig. 135. Now, the evolution of the drive current during pulse is plotted. For long $L_G$ devices, a small decay in

the drive current (<0.4%) can be observed, whereas a constant drive current is observed for $L_G$ = 28nm devices, relevant for this technology. We can thus conclude that we cannot accurately the self-heating effect of our planar of FinFET devices with this pulsed-IV technique.

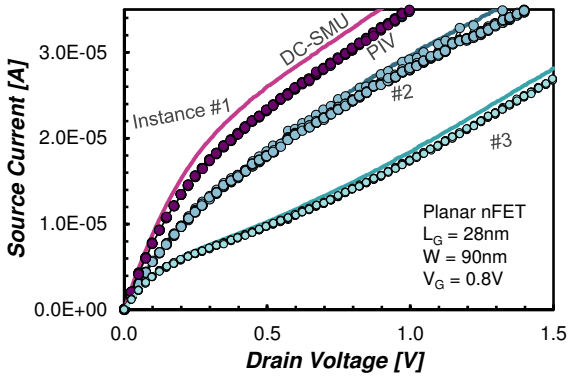



**Fig. 134: Various instances on a device show that PIV measurements exhibit either *similar or lower* drive currents than DC measurements. Note the charge trapping and de-trapping leading to two distinct $I_D$-$V_G$'s for instance #2.**

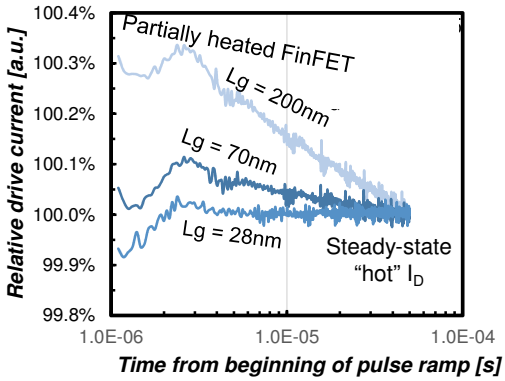



**Fig. 135: A waveform measurement of the pulsed IV-tool on pFinFET devices, depicts the current at the top part of the pulse. For $L_G$=200nm devices, a decay of the current can be observed, whereas this decay is no longer visible for $L_G$=28nm FinFETs.**

### 5.2.2 RF-measurements

In a next step, RF-measurements are performed on planar devices, following the methodology described in Chapter 2. The resulting $g_{ds}$-*frequency* trend, obtained after de-embedding the $Y_{22}$ parameters, is depicted in Fig. 135.

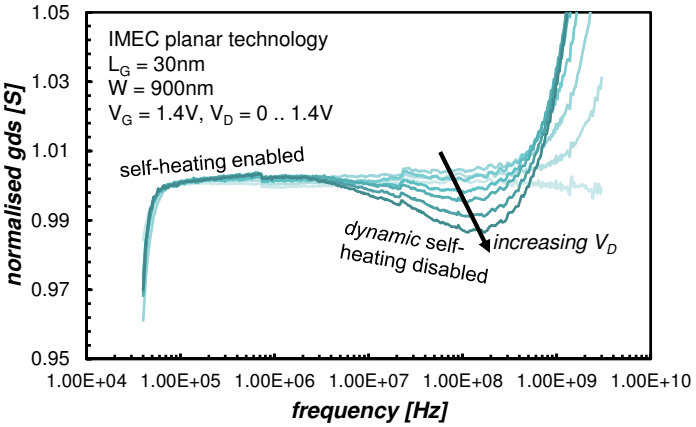



**Fig. 136: RF-measurements show a *drop* at elevated frequency when the dynamic self-heating is disabled. Gate resistance effects, showing as a distinct increase in the $g_{ds}$ are jeopardizing signal analysis at higher frequencies.**

The $g_{ds}$-*frequency* trend shows a *decrease* in conductance at higher frequencies, with a minimum around 200 MHz. Typically, a clear *elevated* plateau is visible at isothermal frequency [Makovejef11].

In these planar devices however, a *decrease with increasing frequency* is visible, which is moreover blurred by the gate resistance, which is a source of an artificial increase of $g_{ds}$ [Makovjev11] also interacts in the relevant frequency range.

If the thermal resistance is to be extracted from these devices, the isothermal frequency has to be chosen from this arbitrary minimum of the $g_{ds}$, which is formed by the self-heating induced $g_{ds}$ decrease and the gate resistance-induced $g_{ds}$ increase. It becomes also clear that the PIV measurements, which

have at most an equivalent frequency of about 5 MHz are too slow to capture the entire effect.

### 5.2.3 Reverse temperature dependence in planar devices

Studying the DC temperature characteristics of our devices, it becomes clear that the expected drive current drop remains absent and that a reverse temperature dependence is observed over the entire measurement window, both in linear as in saturation regime (Fig. 137). This explains the observed drop in conductance at elevated frequencies in the RF-measurements.

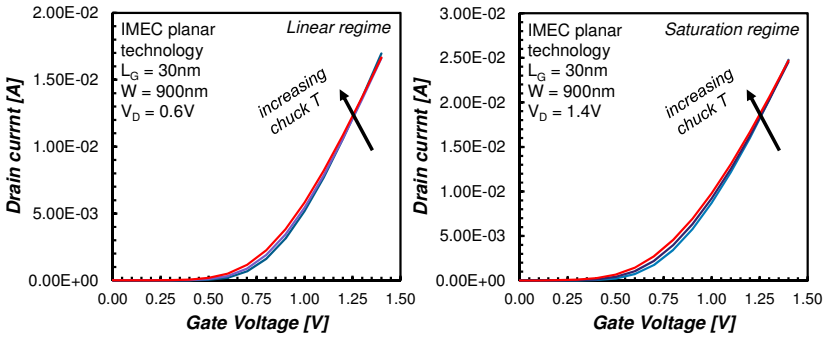



**Fig. 137: Linear and saturation drive current characteristics of a short-channel planar transistor show a lack of drive current decrease at elevated temperatures, i.e. a *reverse* temperature dependence over the entire measurement range.**

Wolpert *et al*. predicted that $dV_{TH}/dT$ would increase for scaled MOSFET nodes, i.e. tending towards a reverse temperature dependence, due to the introduction of high-k dielectrics and metal gates [Wolpert11].

In Fig. 138, the temperature dependence of the main device parameters, fitted with the EKV-model, is depicted [Enz95]. Remarkably, for the studied devices, the $|dV_{TH}/dT|$ is *decreasing* with scaling gate length. This is thus not the origin for the reverse temperature dependence. It is, however, the lack of temperature dependence in the mobility and the strong increase in subthreshold swing which are contributing to the observed reverse temperature dependence.

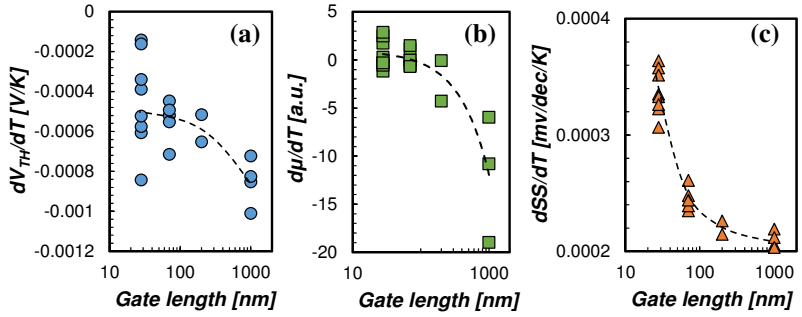



**Fig. 138 (a) The $|dV_{TH}/dT|$ decreases with lower $L_G$, which would tend to yield a more normal temperature dependence for lower $L_G$. However, both the (b) $|d\mu/dT|$ and (c) the $dSS/dT$ by decreasing and increasing respectively, each strongly contribute to the observed reverse temperature dependence.**

It is nonetheless clear that these devices do not show the expected increase in drive current and that the mechanism behind this should be better understood. This requires thorough understanding of the underlying mechanisms. Therefore, indirect temporal analysis of the drive current to assess the self-heating effect is not considered as a suitable technique for these nanometer-sized devices. In the Section 5.3, we will introduce a direct measurement technique that can be applied for this purpose.

## 5.3 Developing a new self-heating measurement methodology

Even though numerous other methods to measure the temperature of an operating semiconductor device exist, most of them fail to entirely capture the heating effect in scaled devices as the heating becomes more localized. In this Section, we focus on electrical measurements as they can be easily performed in-line, whereas physical techniques such as scanning nanoprobe—which potentially has the highest spatial resolution—are expensive and require special structures [Lee07] and fall beyond the scope of this work.

### 5.3.1 Newly developed heater-sensor technique

We propose to measure the self-heating effect in planar transistors using a matched-pair-like structure with two FETs and a common source terminal. One of the devices will be considered the 'heater' device, and the temperature increase will be measured in the other device, the 'sensor'. The device is illustrated in Fig. 139.

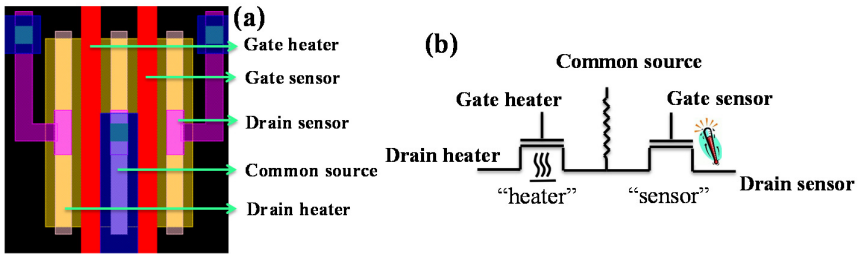



**Fig. 139: (a) Layout of the device and (b) the corresponding schematic representation. The series resistance on the common source node needs to be taken into account in order to keep $V_{DS}$ conditions in the sensor unaltered.**

The temperature can be measured with the sensor by utilizing its temperature dependent parameters. As discussed in the previous Section, the sub-threshold swing (SS) and/or sub-threshold current are strongly temperature dependent parameters and their *sensitivity increases with device scaling*.

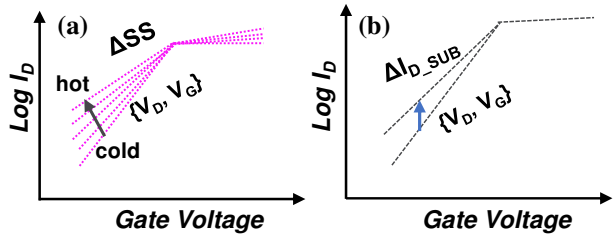



**Fig. 140: Either the (a) subthreshold swing or (b) the subthreshold current can be measured to assess the $\Delta T$.**

An ideal subthreshold swing is typically defined as [Sze81]:

$$SS = \ln(10)\,\phi_T(1 + \frac{C_d}{C_{ox}}) \tag{5.1}$$

with $\phi_T$ the thermal voltage (=kT/q) and $C_d$ the depletion layer capacitance.

The subthreshold leakage current is exponentially dependent on temperature. A common rule of thumb is that the leakage current increases ~2x for every 10K of increase in temperature. The current in this region (i.e. weak inversion) can be described as follows [VanOverstraeten73]:

$$I_D = \frac{W}{L}\mu_0 C_{ox}\left(\frac{kT}{q}\right)^2 \exp\left(\frac{V_{GS} - V_{TH}}{n\phi_T}\right)\left(1 - \exp\left(\frac{-V_{DS}}{\phi_T}\right)\right) \tag{5.2}$$

where $\mu_0$ is the carrier mobility, $W$ and $L$ transistor width and length, and $n$ a non-ideality factor.

First, the temperature dependency of 'sensor' is separately calibrated by ramping the external temperature with a thermo-chuck. The SS typically shows linear behavior over a broad temperature range, with a sensitivity of ~3mV/dec-K for silicon channel devices. Therefore, the sub-threshold swing reduction in the 'sensor' can clearly give information on the heat dissipated in the device and the thermal resistance $R_{TH}$ of the structure. Moreover, the sensor is not prone to any degradation mechanisms since it remains biased in low $V_{GS}$/$V_{DS}$ conditions. To verify this, the end-of-measurement SS of the sensor was compared to the initial SS, showing no degradation, in contrast to the $I_DV_G$ characteristics of the DUT.

Subsequently, the change in SS of the sensor is extracted by fitting the EKV-model of the $I_D$-$V_G$, for varying biasing conditions of the 'heater' FET in strong heating conditions (high $V_D$ and $V_G$).

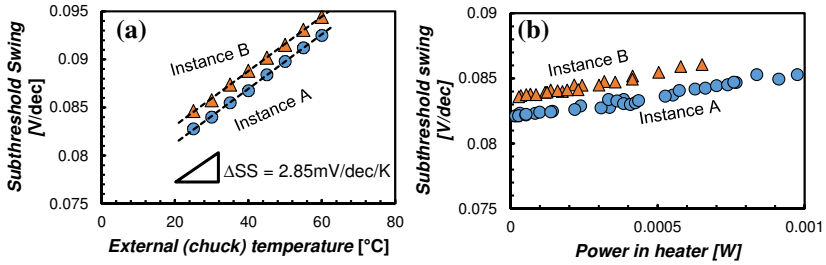



**Fig. 141: (a) Calibration of the subthreshold swing (SS) of two different instances (A and B) of the same device which varies linearly with the externally applied chuck temperature. (b) Change in the subthreshold slope when power is dissipated in the heater.**

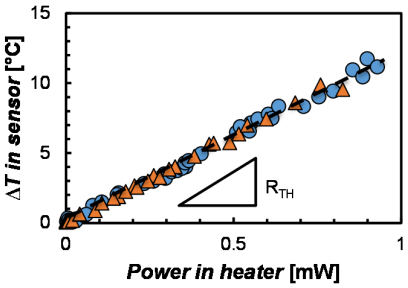



**Fig. 142: The resulting slope of the power vs ΔT equals the thermal resistance $R_{TH}$ of the structure at a certain distance from the heater, and is identical for both instances of the same device.**

A key factor in this structure is the gate-to-gate spacing between heater and sensor. Because we are using a common-source, the distance is limited to only a single poly pitch. Moreover, both devices share the same active area, i.e. there is no shallow-trench-isolation (i.e. lowly conductive $SiO_2$) in between.

Finally, from wafer-level measurements, it also becomes clear that subthreshold swing is the most sensitive parameter to utilize for the heating effect (Fig. 143).

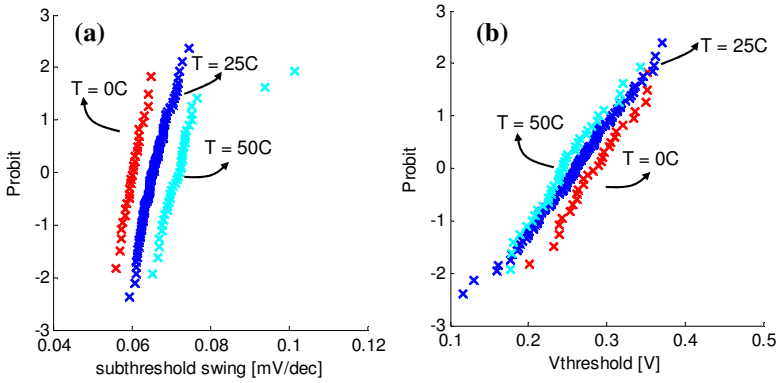



**Fig. 143: (a) Subthreshold distribution versus on temperature shows little variation and clear-cut distributions, whereas (b) the impact on the $V_{TH}$ is much smaller and the distributions are overlapping.**

## *5.3.2 Measurements on planar devices*

In this Section, we analyze the set of devices that are available. They exist of a series of variations in *gate length*, *width* and *separation between gates*. An overview of the studied devices is given in Table IV.

**Table IV: Test-devices in this study**

| Device name | L [nm] | W [nm] | Gate-to-Gate separation | Active Area [nm$^2$] |
|---|---|---|---|---|
| A1 | 50 | 70 | 100 | 2.74E+04 |
| A2 | 80 | 70 | 100 | 3.16E+04 |
| A3 | 50 | 200 | 100 | 7.84E+04 |
| A4 | 80 | 200 | 100 | 9.04E+04 |
| B1 | 50 | 500 | 100 | 1.96E+05 |
| B2 | 50 | 500 | 140 | 2.16E+05 |
| B3 | 80 | 500 | 100 | 2.26E+05 |
| B4 | 80 | 500 | 140 | 2.46E+05 |

The obtained heater-sensor measurements are depicted in Fig. 144. It is clear that for all devices, *even though they are fabricated in planar technology*, a $\Delta T$ induced by self-heating can be observed. At the same

operating conditions, the wider devices show the higher *absolute* temperature increase because they dissipate more power than their narrow counterparts. The devices with longer $L_G$ show lower $\Delta T$. From Fig. 144 (b), we observe an distinct impact of increasing the gate-to-gate separation, i.e. increasing the distance from heater to sensor.

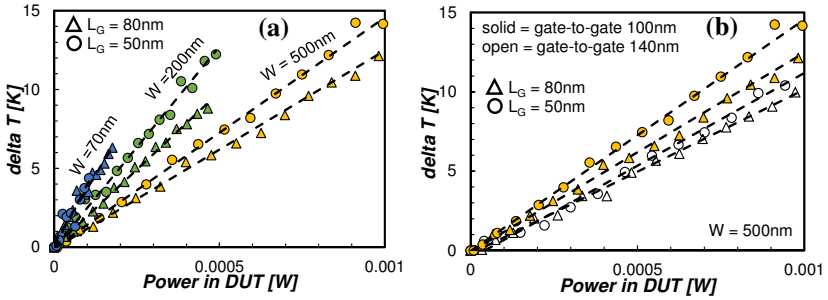



**Fig. 144: (a) The experimentally obtained data for self-heating shows that all devices show linearly increasing temperature with power and (b) increasing the heater-to-sensor distance reduces the sensitivity of the sensor.**

In Fig. 145 the thermal resistances $R_{TH}$ are extracted for all the devices. In this case, the thermal resistance is directly extracted from the ratio of $\Delta T/Q_{device}$ (as discussed in Chapter 2), which is typically constant over the entire measurement range. The non-normalized $R_{TH}$ for the narrow devices strongly increases over the wide devices (The $R_{TH}$ is not normalized to the width of the device). This can be expected as less volume of Si is available in a narrow device per unit of heat generation, due to the smaller footprint of the device.

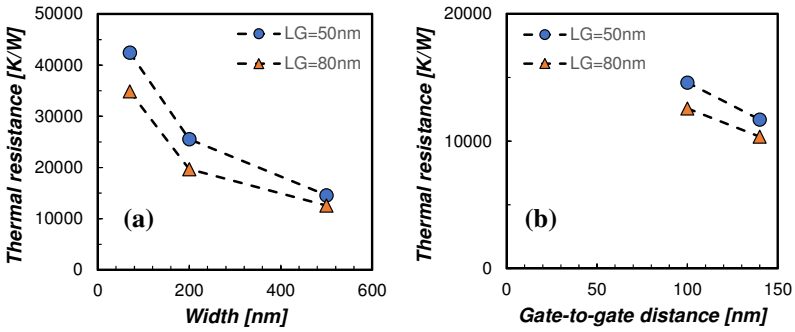



**Fig. 145: The thermal resistance clearly increases for the narrower devices, i.e. for the *same amount of dissipated heat*, the narrow devices will heat more, whereas increasing the gate length clearly mitigates the self-heating effect. (b) Increasing the distance between the heater and sensor shows to reduce the sensitivity of the sensor.**

When the $R_{TH}$ is normalized to the width of the device, i.e. taking into account that *a device will produce heat proportional to its width*, it appears that the narrow devices have a *lower* normalized $R_{TH}$ than the wide devices (Fig. 146). This can be understood by the fact that in a narrow device, the heat can escape in all dimensions, whereas in a wide device, the heat flux becomes a two-dimensional case with no lateral heat transfer.

Taking both of the above conclusions into consideration, it is clear that for FinFET devices, which will have a higher drive current at the same footprint, but will still consist of multiple parallel fins, the self-heating effect will increase.

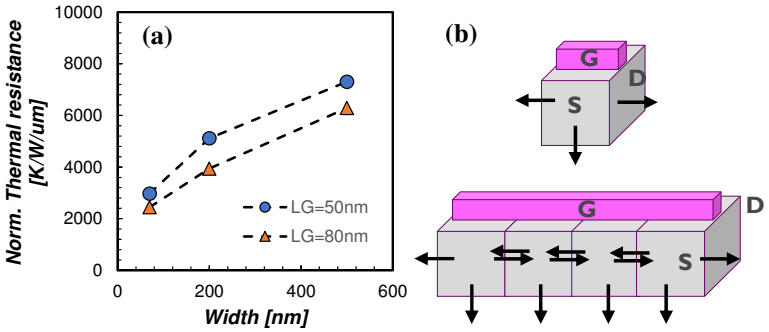



**Fig. 146: (a) The $R_{TH}$ normalized to the width of the device decreases for narrower devices, because a narrow device is more effective in releasing its the heat since its lateral surroundings are cold, illustrated in (b).**

### *5.3.3 Extracting the series resistance*

An additional advantage of the above described heater-sensor structure is that it allows to directly measure the series resistance on the common source node. The composite series resistance of the source junction and contact will give rise to a bias increase at the source junction. This will result in a net current between source and drain of the sensor, even though a virtual $V_{DS}$ = 0V is applied on the sensor. By forcing zero current through the drain node of the sensor, the actual voltage of the source junction can be determined. The secondary drain in this structure will thus act as a Kelvin-type measurement. The principle is illustrated in Fig. 147.

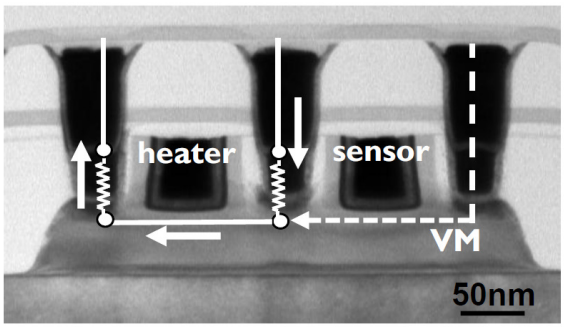



**Fig. 147: TEM image of the heater-sensor structure and illustration of how the series resistance of the common source junction+contact can be extracted by utilizing the drain of the sensor as voltage-monitor.**

If the series resistance is non-negligible, it will result in certain additional power dissipation in the contact, in addition to the power dissipated in the channel. Fig. 148 shows the channel resistance and the relative size of the series resistance on the source junction and contact. Around $V_{DD}$ conditions, this series resistance can contribute up to 20% of the total resistance between source and drain. This means that also this amount of heat will be generated in the contact, rather than in the channel, and should therefore be taken into account in simulations.

An application of this measurement technique will be given in Section 5.6.3, where we will assess the difference in the junction resistance for nFET and pFET and evaluate its impact on the accuracy of the heater-sensor measurement. Meanwhile, we propose this structure as a valuable alternative for typical indirect series resistance estimation techniques for MOSFET characterizion, for example by gate length regression, which typically suffers from extrinsic effects, such as non-constant doping.

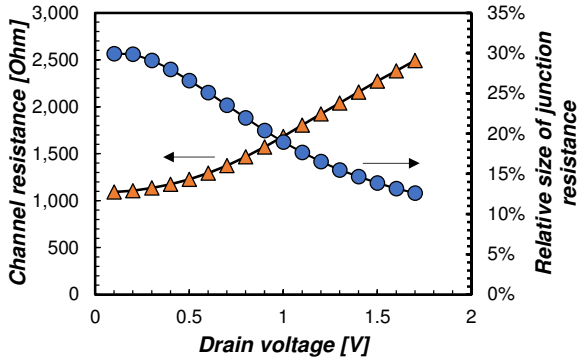



**Fig. 148: The composite resistances of source junction and contact can be extracted separately from the channel resistance. Around typical $V_{DD}$ conditions, about 20% of the power will be dissipated in the source contact.**

## 5.4 Self-heating simulations

As the main interest is the actual temperature *of the DUT*, the result of the heater-sensor measurements discussed in the previous paragraph needs to be corroborated with simulations. The simulations can facilitate extracting the temperature profile in the structure, and the average and peak temperatures in the channel of the 'heater' and 'sensor' device. In this Section, we will apply several simulation techniques and discuss the results and the merits of each of the techniques.

### *5.4.1 Classic 3DFEM simulations*

For the 3DFM simulations, the geometry of the device is rebuild based on TEM data and information on the layout, illustrated in Fig. 149. Subsequently, a heat source is inserted in a thin volume under the 'heater' gate, representing the channel. The input power can be extracted from the measurements or a unity power density can be used for thermal resistance extraction as this is a linear system. At this point, also the series resistance of the source junction is taken into account, which contributes to ~20% of Joule heating in the device

in $V_{DD}$ conditions, as discussed in the previous Section. Subsequently, Fourier's law of heat diffusion is solved:

$$q = -\kappa \nabla T \tag{5.3}$$

where $q$ is the volumetric energy addition or heat flux (W/m$^3$) and $\kappa$ thermal conductivity (W/m.K).

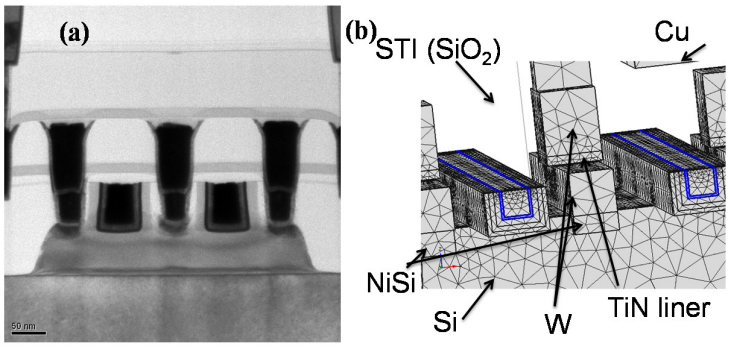



**Fig. 149: (a) TEM micrograph of the heater-sensor transistor pair and (b) 3DFEM model of the device.**

Dirichlet boundary conditions are applied at the outer edges of the BEOL, representing the chip operating at room temperature.

Another important aspect of these simulations are the *thermal conductivity parameters* of the materials utilized for processing these devices. Even though bulk thermal conductivities for the studied materials are widely available, their values can strongly vary due to geometric confinement or temperature, due to phonon scattering mechanisms. An example of this is given in Fig. 150, showing that a 100nm film of Si has ~2x lower thermal conductivity than bulk Si [Liu06].

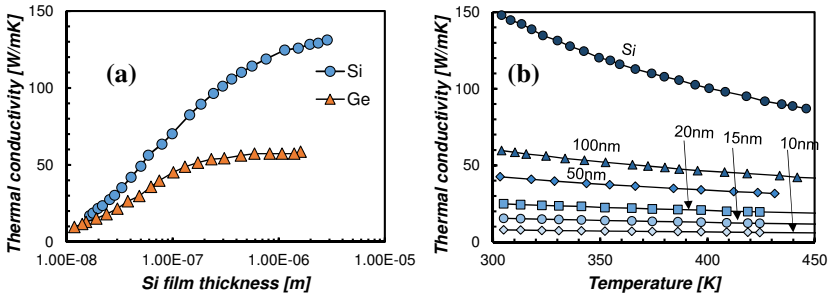



**Fig. 150. Thermal conductivity of thin-film silicon is strongly dependent on thickness and temperature [replotted from Liu06]**

Moreover, as semiconductor films approach nm-size, their thermal characteristics become anisotropic [Aksamija], [Quintana11]. This effect will be taken into account in Chapter 6 by the use of thermal conductivity *tensors*. For the simulations below, thermal parameters for other materials are extracted from literature sources, and their thickness specific value is utilized depending on their occurrence. For silicon, the temperature dependence is taken into account. Their values are described in Table V.

**Table V: Thin-film thermal conductivity values for the materials occurring in the DUTs**

| Material | Occurrence | κ [$WK^{-1}m^{-1}$] | Reference |
|---|---|---|---|
| Si | Substrate | 120 | [Liu13] |
| | Active Box (Plan.) | 90 | |
| | 20nm film (Fin) | 40 | |
| | 10nm film (Fin) | 20 | |
| $SiO_2$ | STI | 1.2 | [Liu06] |
| NiSi | S/D contacts | 16 | [Deng97] |
| W | IM1, IM2, VIA1 | 40 | [Lu09] |
| Cu | Metal1 | 250 | [Lu09] |
| HfO2 | Gate oxide | 0.5 | [Panzer09] |
| TiN | Gate liner | 4 | [Kim11] |
| TaN | Gate liner | 3.4 | [Grayeli11] |
| Al | Gate metal fill | 21 | Stojanovic[11] |

An illustration of the extracted heat profile is given in Fig. 151. From these simulations, it is clear that the sensor will only capture a fraction of the $\Delta T$ in the DUT. In the case of a 500nm device, the average temperature in the sensor will be 2.18x lower than the *actual* average temperature in the heater. For the narrowest devices, this ratio increases up to 3.57x. This indicates that the sensitivity of the sensor is actually decreasing as the devices are getting narrower. This can be explained by the narrower silicon volume available for heat transfer along the device and the reduced intrinsic thermal conductivity of the silicon.

The heat flux vectors indicate that the main heat dissipation path is through the bulk. Only a fraction of the heat is dissipated along the contacts and the gate.

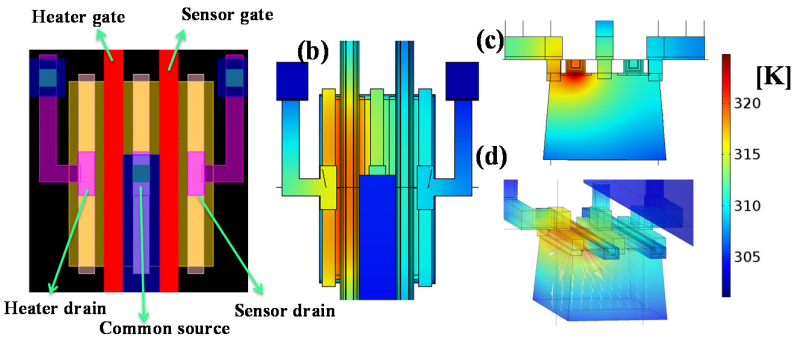



**Fig. 151. (a) Top level overview of the heater-sensor structure. (b-d) Illustration of the heat distribution in the heater and the sensor by finite-element-simulations at $V_{DD}$ = 1.5V. In (d) also the heat flux vectors are depicted.**

The simulated values for the sensor can be compared against the measured values, for both typical ($V_{DD}$ = 1V) and high $V_{DD}$ conditions ($V_{DD}$=1.5V), yielding power densities of ≈0.3 and ≈1.7mW/um respectively. The results are depicted in Fig. 152 and show a reasonably good match. At same operating conditions, the sensors in narrow devices measure a lower $\Delta T$ than their wide counterparts, the sensors become thus *less sensitive as the device width decreases*.

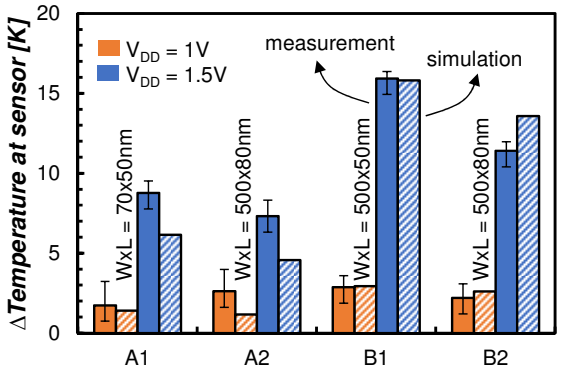



**Fig. 152: Measured versus simulated values for the sensor temperature for various instances of the DUT, simulated for the medium and high $V_{DD}$ conditions, yielding power densities of 0.3 and 1.7 mW/um respectively.**

Finally, the average channel temperature of the DUT (i.e. the heater) itself can be plotted (Fig. 153). Narrowing the device (e.g. from type B1 to A1) has a slighter effect on the heater FET (~20% lower heater $\Delta T$ simulated for A1 than B1) than what is observed on the sensor-side (~50% reduction in sensor $\Delta T$) in Fig. 152.

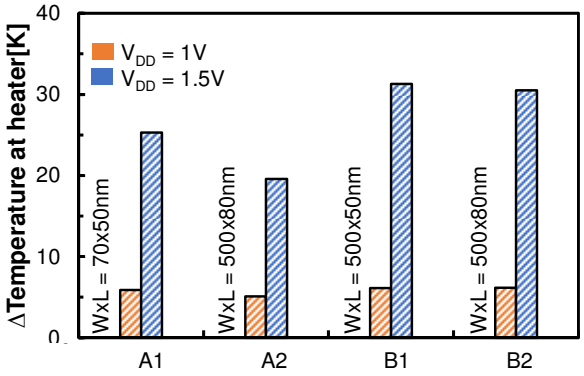



**Fig. 153: Simulated ΔT in the DUT for various instances for the two $V_{DD}$ conditions.**

These simulations also allow to make an estimation of the time-constants of the SHE by solving the time-dependent Fourier equation:

$$\rho C_p \frac{\partial T}{\partial t} = \nabla \cdot (\mathrm{k}\nabla T) + q \tag{5.4}$$

where $\rho$ the density of the material (kg/m$^3$), and $C_p$ the specific heat (J/kg.K) of the material. The outcome of these simulations thus not only hinges on the correctness of the thermal conductivity, but also of the specific heat. In this case—and since there is no direct reason to assume otherwise—bulk values for the specific heat are utilized. The outcome of the time-dependent simulations is given in Fig. 154.

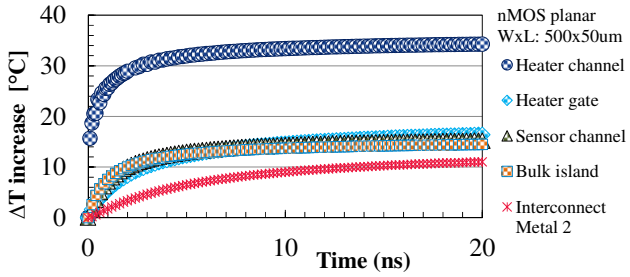



**Fig. 154. Time-dependent solution for the self-heating effect is simulated for structure B1. The device channel heats up within a few ns, whereas the steady state for the BEOL is not yet fully reached within ~20ns.**

The time-dependent simulations reveal the thermal time constants for heating the channel and confirm the above failure to observe self-heating in the PIV measurements . The channel will reach its steady-state temperature within a few ns, the back-end interconnect metal is heated within ~20ns. A time constant of the channel (i.e. corresponding to roughly the 63% value) of about 1ns ~1GHz is obtained for the channel ΔT. This is in line with the RF-measurements, where the isothermal plateau was blurred by the gate resistance effects (i.e. we can only provide a lower-limit), and where a minimal $g_{ds}$ was found at ~200MHz.

### 5.4.2 Particle-based electro-thermal simulations

To capture the full heat profile in the device and to analyze the effect on the drive current, an electro-thermal (ET) simulator is required. When using particle-based ET simulations, we are moving away from the commonly used Joule heating model used in commercial device simulators.

Because the above simulations more accurately represent the optical phonon to acoustic phonon bottleneck, they give rise to more pronounced hot-spots. Next to a 3DFEM simulation, a 2D-simulation with ET-MC simulator is depicted in Fig. 155, which is the limiting case of an infinitely wide transistor, since the heat can only flow in plane. The $V_{DD}$ conditions in the simulations are increased up to 1.66V, to obtain the right total power dissipation in this device. In strong contrast to the obtained measurements and de 3DFEM simulations, however, the sensor region shows barely any temperature increase. For both simulations, the same boundary conditions were applied.

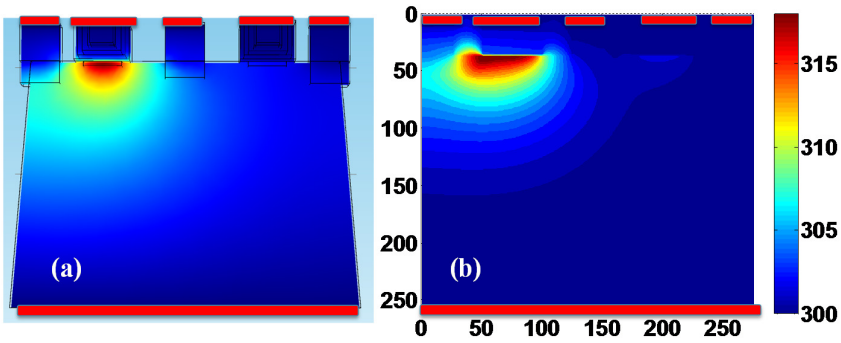



**Fig. 155: (a) Average lattice temperature profile in the active silicon layer obtained with 3DFEM simulations and (b) obtained with the thermal Monte Carlo device solver, indicating a pronounced hotspot in the channel near the drain-side of the heater, but very little temperature increase in the sensor region. Note that both simulations utilize the same boundary conditions, i.e. nearby the FET contacts, indicated with the red bars.**

### 5.4.3 Impact of boundary conditions: multi-scale approach

A discrepancy between the results of the 3DFEM in Section 5.4.2 (Fig. 155) can be observed with the earlier results in Section 5.4.1 (Fig. 151). This can be fully attributed to the boundary conditions. It is clear that the obtained temperature profile is strongly influenced by the presence of Dirichlet boundary conditions at the contacts. Therefore, the EC-MC simulations will fail to extract the effective thermal resistance of the device.

Ideally, the full BEOL is co-integrated in the ET-MC simulations. However, the particle-based simulations are computationally intensive and time-consuming. We propose therefore a hybrid multi-scale solution, combining the FEM simulations (for the BEOL) with ET-MC simulator. Practically, this can be implemented by forcing the thermal boundary conditions obtained with the 3DFEM on the 2D solver. The result of this is shown in Fig. 156. Now, the ET-MC simulator shows the exact heat profile and the results are quantitatively consistent with the measured temperatures in the sensor.

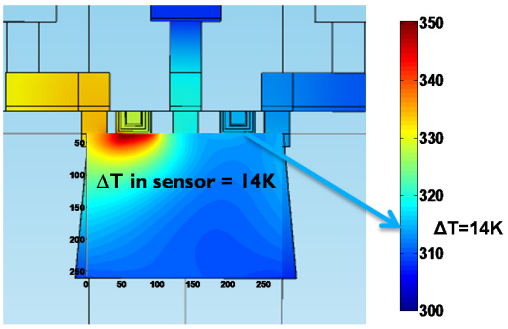



**Fig. 156. Combining 3D FEM simulations with an ET-MC simulator will yield the exact heat profile and quantitatively correct temperatures for the entire structure. In this example device with 2nm EOT is simulated for a $V_G = V_D = 1.66$V. The resulting sensor $\Delta T$ is 14K.**

The concept of co-integrating multiple simulation tools was further elaborated by Qazi *et al*, where the *Giga3D* module within Silvaco Atlas was used to simulate the device interconnect level, which provides the temperature boundary conditions at the device level for the ET device simulator. A

MATLAB framework was then used to interface these two different simulation modules together [Qazi15].

### 5.4.4 Proof-of-concept: common source versus common drain

As it could be observed from the ET simulations in the previous paragraphs, the hot-spot of the device appears to be located at the drain, because the electrons obtain their highest drift velocity and thus energy at this location. This is unfavorable as the heater's drain node is located further apart from the sensor than the (common) source node. If the DUTs can be biased in a common drain configuration, the hotspot could be located closer to the center and thus increase the sensor's efficiency.

This can be obtained by changing the bias conditions of the DUT. In a *common-drain* configuration, $V_{DS}$ of the heater is dynamically adjusted, thereby necessitating corresponding adjustments in $V_G V_S$ and $V_B$ of the *sensor* throughout the sweep, to avoid unwanted changes in the sensor's potential distribution. However, to avoid forward-biasing the junction, these measurements can only be performed in a limited bias range.

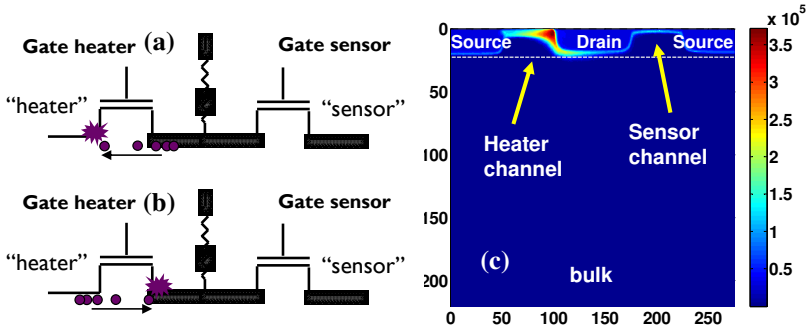



**Fig. 157: Switching the structure from (a) common-source to (b) common-drain can bring the center of the heat source closer to the sensor as (c) the electron drift velocity (m/s) reaches its maximum at the drain side.**

The measurements in Fig. 158 indeed show that in a common drain configuration, the thermal resistance as observed with the sensor is strongly increased (from 15 to 24 K/mW, i.e. 60% increase in the observed $R_{TH}$), thereby experimentally proving the existence of the hotspot near the drain.

The noise in the data however increases significantly. We attribute this increased noise due to the harsh operating conditions of the sensor, which endures a large gate to bulk bias during the common drain measurement.

The measurements are corroborated with the above described multi-scale ET simulations in Fig. 159. The simulations show a good quantitative match: the $R_{TH}$ increases 42% when changing from common source to a common drain configuration.

We can thus conclude that the heat distribution in the DUT is strongly non-uniform and that a hotspot is formed at the drain-side of the device. Moreover, we showed and that we can shift the position from the hotspot by changing the bias conditions.

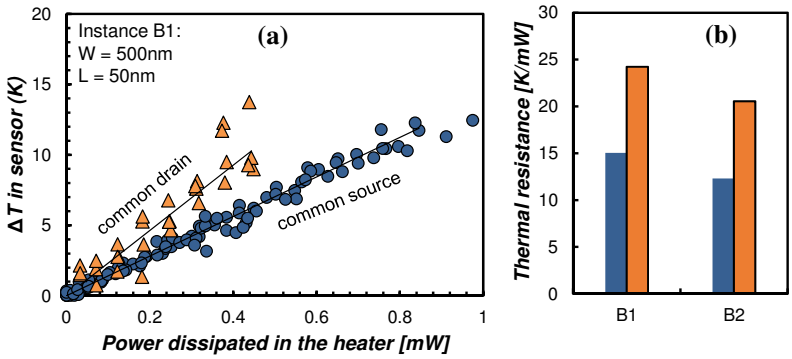



**Fig. 158: (a) Measurement results on multiple instances in switching from common-source to common-drain setup for the widest device. (b) Extracted thermal resistances show higher sensitivity in the common-drain configuration.**

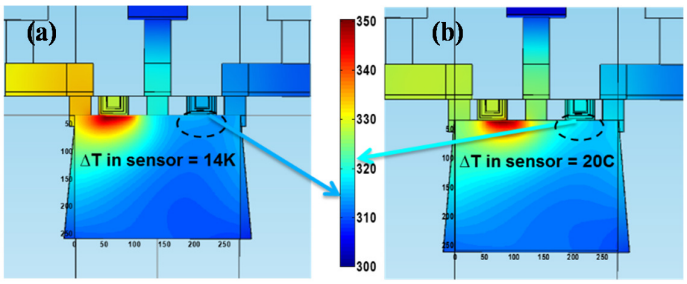



**Fig. 159: Electro-thermal simulations of the (a) common source versus (b) common-drain configuration showing the clear change of the hotspot location closer towards the sensor DUT.**

### *5.4.5 Conclusions*

In the above Sections, we have presented a sound measurement technique to assess the self-heating effect in nanoscale devices. We have corroborated the technique with various simulations methodologies. From the simulations, it became also clear that this test-structure will lose its sensitivity when the device width is scaled, as relatively less heat will reach the sensor.

In the Section 5.5, we apply and corroborate the measurement techniques in a case study.

## 5.5 Case study: Assessing temperature effects on FinFET and GAA-NW devices

In this case study, predictive simulations to compare planar with FinFET devices are performed. Modifications to the test-devices enable to repeat the planar experiment for FinFET devices. The measurements of the FinFETs are benchmarked against the simulations. Subsequently, we experimentally assess the impact of FinFET topology modifications on the SHE and compare the results with imec's first generation of stacked GAA-NWs. Finally, we discuss the impact of SHE on the time-zero performance and reliability.

### 5.5.1 Converting from planar to FinFET

In this Section, we simulate the self-heating effect in the heater-sensor device for FinFETs. A partial revision of the maskset allowed a conversion of the planar devices to FinFET devices (Fig. 160). Moreover, a new device was introduced where the fins of the heater are disconnected from the fins in the sensor device. This signifies that the heat between the heater and the sensor will predominantly flow through gate lines and the common source. A list of devices is given in Table VI.

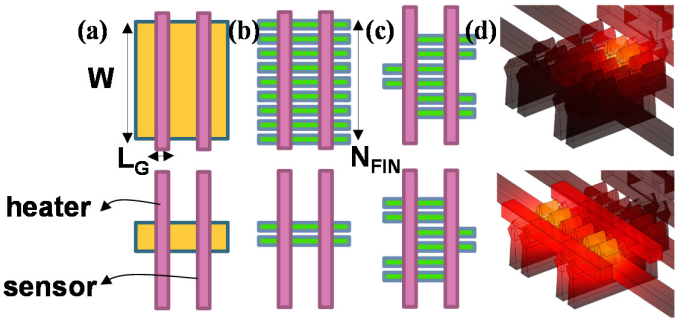



**Fig. 160. (a) Heater-sensor structures in planar devices, (b) converted to FinFET with various $N_{FIN}$. (c) An additional test-structure for FinFETs with separate heater and sensor fins, i.e. only connected by the common source and the gates. (d) Example of the heat flow in the misaligned structures.**

After converting the planar devices to FinFET structures, we will now measure and simulate all the proposed device structures from Table VI. Measurement results are depicted in Fig. 161 for both $N_{FIN}$=10, $N_{FIN}$=2 and the aligned versus misaligned cases. To compare measurements and simulations with the planar devices, the following normalization is proposed: the equivalent "planar" width per fin corresponds to 70nm, taking into account the fin width and the sidewall height (i.e. $W_{EQUIVALENT} = W_{FIN} + 2 \times H_{SIDEWALL}$). This means that the *power density* obtained for planar devices of 1.7 W/um can be translated into an equivalent *power per fin* of 1.2 x $10^{-4}$ W/fin.

**Table VI: List of FinFET devices used in this experiment.**

| Device name | L [nm] | Heater (Nfin) | Sensor (Nfin) | Gate-to-gate seperation | Aligned |
|---|---|---|---|---|---|
| A1 | 50 | 2 | 2 | 100 | Yes |
| A2-I | 50 | 4 | 2 | 100 | No |
| A2-II | 50 | 2 | 4 | 100 | No |
| A3 | 80 | 2 | 2 | 100 | Yes |
| A4-I | 80 | 4 | 2 | 100 | No |
| A4-II | 80 | 2 | 4 | 100 | No |
| B1 | 50 | 10 | 10 | 100 | Yes |
| B2 | 80 | 10 | 10 | 100 | Yes |
| B3 | 50 | 2 | 2 | 140 | Yes |
| B4 | 80 | 2 | 2 | 140 | Yes |

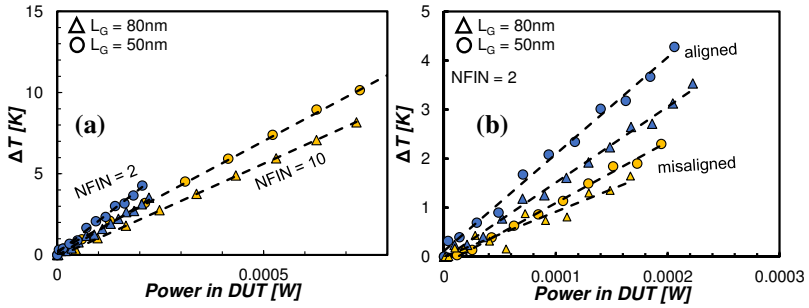



**Fig. 161. Measurements for the various cases according to Table VI, (a) 2 fin versus 10 fin devices and (b) aligned versus misaligned fins.**

The FinFETs show strong dependence of the *device geometry* on the (absolute) temperature measurements (Fig. 161) in which the highest absolute temperatures (at identical operating conditions) can be measured with the shortest $L_G$ and the highest $N_{FIN}$. For that particular case and a power density of $1.2 \times 10^{-4}$ W/fin, this results in a *sensor* $\Delta T$ of 15K and an *actual* $\Delta T$ of the *heater* of 91K.

The variations on the actual heater $\Delta T$, obtained from the simulations are much smaller. This means that the *sensor temperature approximates the heater temperature better as the gate length is decreasing and the number of*

*fins increases*. Therefore, for benchmarking in Section 5.6, we focus the devices of type '*B1*'.

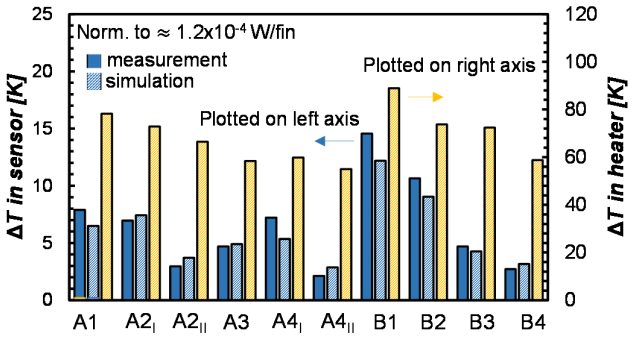



**Fig. 162. Simulation results for the various cases according to Table VI, normalized to identical power densities per fin. The highest ΔT is found in devices with 10 fins (B1), making it the best 'sensor' for further use.**

Fig. 162 shows the simulation results for the various cases according to Table VI. To compare the measurements and the simulations, only the sensor $\Delta T$'s can be compared. A reasonable match between the simulation data is found. Also the particular case of the misaligned devices are matching the measurements results well. This indicates that the thermal conductivities obtained from literature for the respective gate and the local interconnect material are reliable and that we can utilize them to make projections for more scaled device nodes in Section 5.7.

### *5.5.2 Comparing planar FET versus FinFET*

Subsequently, we compare the obtained simulation results for FinFETs and planar devices in Fig. 163. For the planar devices, our simulations indicate that the channel temperature increases from 25 up to 34K depending on the width of the devices, whereas for FinFETs the simulations indicate that ΔT can increase from 78K for a 2-fin device, up to 88K for a 10-fin device (Fig. 163). The increased heating in FinFETs is not only due to the lower heat conducting volume, but also due to the decrease in thermal conductance originating from phonon boundary scattering.

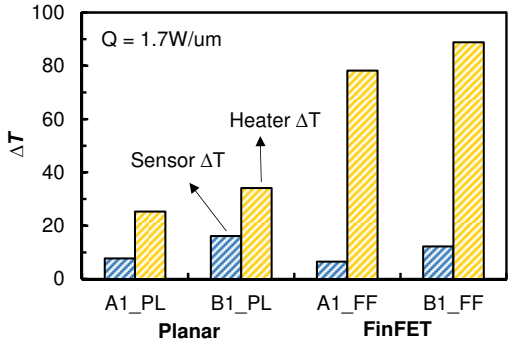



**Fig. 163: The amount of captured heat in the sensor is similar for planar as for FinFET devices, whereas the actual temperature in the channel will be much higher in the FinFET.**

These effects also have the consequence that *the sensor will capture relatively much less heat from the heating FinFET* as it could for a heating planar device because the heat stays more confined in and around the DUT and is less spread out towards the sensor. Still the heater-sensor structure can be used as a good relative benchmarking tool *within one technology,* in other words, if sensitivity of the sensor is not changing between the devices i.e. as long as the devices are mostly identical in their geometry.

The high temperature increases in FinFETs could potentially impact the device's vulnerability to temperature-activated degradation mechanisms in the gate stack, such as BTI and HC degradation. Therefore, we will discuss these effects in Section 5.8.2.

## 5.6 FinFET technology benchmarking

In this Section, we benchmark various variations and optimizations of the FinFET technology in terms of source/drain epitaxial growth for reducing of series resistance, study fin height variations and finally investigate the impact of the device polarity on the self-heating effect.

### 5.6.1 S/D epi variations

Due to the 3D structure of the FinFET body, the shrunk dimensions of the source and drain contacts result in a large S/D series resistance. For this reason, diamond-shaped S/D epi structures were proposed to reduce this parasitic resistance [Kawaski09]. As a result, the S/D exhibits a slightly extended width with respect to the channel. Multiple integration options have been proposed to generate these diamond shaped S/D epi structures. One of the first was the 'raised' S/D, in which Si on SiGe was grown on top of the original fin, along the preferential crystallographic planes. Another option is the embedded S/D, in which a part of the fin is recessed prior to S/D growth. The latter avoids the presence of crystallographic defect planes and allows additional S/D induced stress in the channel, specifically in the interest of pFET devices (to increase the mobility) with SiGe contacts. Illustrations and TEM micrographs of both techniques are given in Fig. 164.

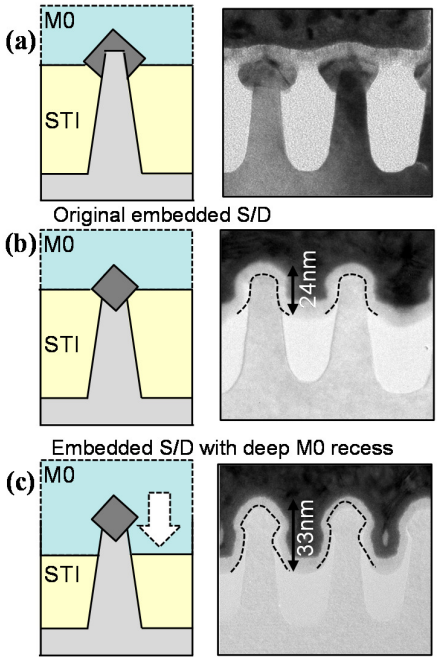



**Fig. 164: (a) Raised and (b) embedded integration of the source/drain contact. (c) TEM micrographs of raised EPI, embedded EPI and embedded EPI with more wrap-around of the local interconnect respectively.**

The result of the benchmark for the various EPI geometries (raised versus embedded) are depicted in Fig. 165. The embedded EPI indicated as (b) in the Fig. 164, shows a distinct increase in the $R_{TH}$ over both the old raised and the newest embedded EPI implementations. This increase can be attributed to the recess depth of the local interconnect metal, which result in a wrap-around the S/D contacts, which is 40% smaller in this case. Modifying the P dopant content in the most deeply recessed S/D EPI devices initially showed to increase the $R_{TH}$ over the non-doped S/D EPI (impurity scattering is known to reduce the phonon-mean-free path), but this effect was not confirmed in more recent devices where the thermal resistance was reduced similar to the value of the pure Si S/D.

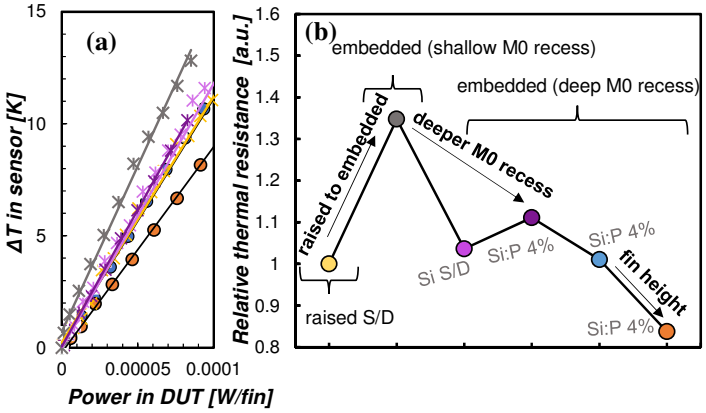



**Fig. 165: (a) Benchmark results of thermal resistance of FinFET (b) Raised EPI (yellow) and the embedded EPI with deeper M0 recess show (blue and purple) lowest thermal resistance. Fin height modulation reduces $R_{TH}$ further(see next Section).**

## 5.6.2 *Fin height variations*

In this section, we study the impact of FinFET height scaling on the thermal resistance. We increase the effective fin height *by performing a stronger recess of the gate over the fin*. This delivers a higher effective width of the device for the same footprint area. Illustration of the experiment and the results are depicted in Fig. 166.

From the experiment, we observe that the taller fins result in an increase of 24% of the saturation drive current. Moreover, the thermal resistance of these taller devices is *reduced*, because the average distance from the heat source (~ equivalent to the mass center of an object) is closer to the bulk substrate, which is the main heat dissipation path for silicon FinFETs.

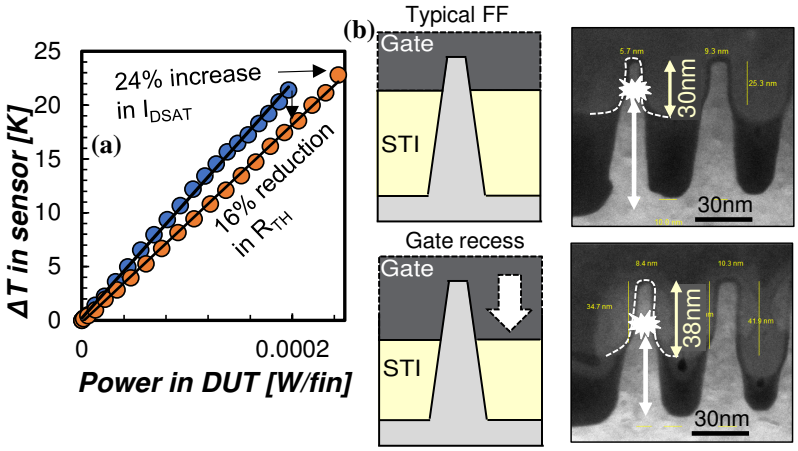



**Fig. 166: (a) Impact of FinFET height scaling on the drive current and thermal resistance. Note that the overall heat generation for the taller FET will be higher. (b) Schematic and TEMs of both FinFET geometries, showing the mass center of heat being closer to the bulk for the tallest fin.**

### *5.6.3 nFET vs pFET*

The DoE's for the above described devices only consists of nFET polarity devices. In unichannel processed wafers, the n-type devices will become p-type. This enables us to compare between nFET and pFET devices. We do not expect an *intrinsic* difference expected in thermal characteristics between nFET and pFET devices if only the channel and the junction doping is adapted.

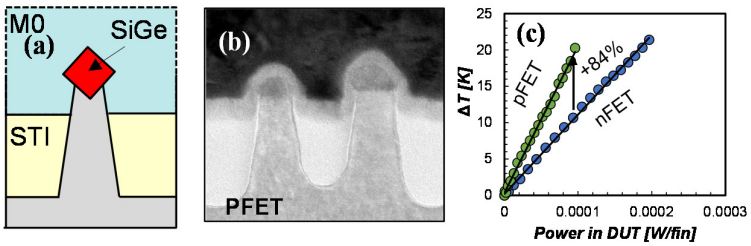



**Fig. 167: (a) Illustration and (b) TEM of pFinFETs with SiGe present at the S/D. (c) Impact of FinFET polarity on the thermal resistance.**

In Fig. 167, we illustrate the devices and show a TEM, and show the result of the measurements. In this case, the source/drain EPI for the pFET devices consists of *SiGe* instead of Si:P for nFET devices. This has drastic consequences on the thermal conductivity of the devices. The thermal resistance of the pFET appears 84% higher than its nFET counterpart. Given the fact that only a fraction of heat dissipates through the contacts, the *difference is much larger than expected*.

In Fig. 168, we show that the *parasitic series resistance of the pFET* is influencing this result, by generating additional heat in common source contact. The shared source junction is located directly next to the sensor device. Heating of the common source will therefore have an effect of delocalizing the heat source position, and thus have a large impact on the measured temperature in the sensor FET.

This is confirmed by electrical measurements of the common source contact resistance which was extracted using the sensor's drain as voltage-monitor. In the pFET case, the contact resistance is up to ~13x higher w.r.t. nFET, which remains in the order of 100 Ohm in any other case. The pFETs contact resistance becomes thus significant w.r.t. the channel resistance and effect on the sensor $\Delta T$ is confirmed by simulations. The series resistance on the common contact has an influence both on the temperature on the device itself, but (relatively) much stronger on the sensor.

In conclusion, we thus cannot make any quantitative statement on the increase of the thermal resistance in nFET devices over pFET devices because of the parasitic series resistance effect.

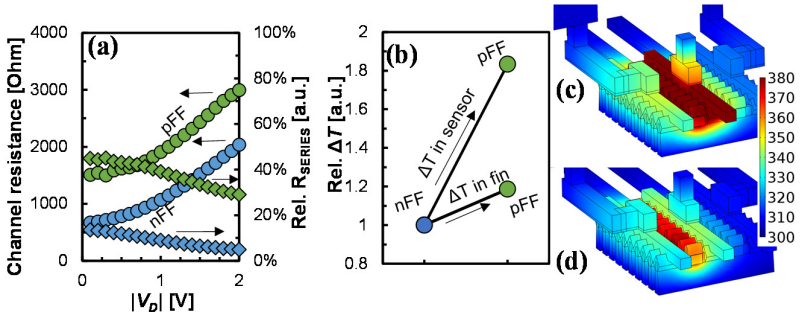



**Fig. 168: (a) Measured channel resistance for a nFF and pFF and the relative contribution of the series resistance a.f.o. $V_D$ (b) Simulated ΔT in the sensor and the DUT for nFF and pFF taking into account the series resistance (and thus heating) in the common source contact. (c) Illustration of the heating of the common source contract in a 10-fin device, (d) which remains cold in the nFF case.**

### *5.6.4 Experimental: GAA-NW devices*

In this last Section, we compare the thermal resistance of FinFET and stacked GAA-NW devices [Mertens16] which have similar drive current per unit footprint. An illustration and TEM of the devices is shown in Fig. 169.

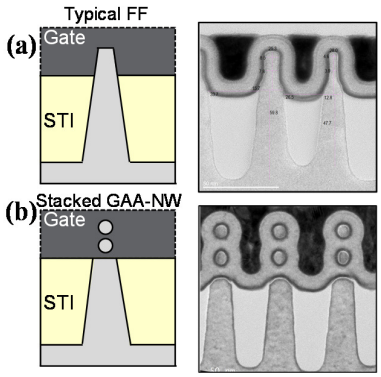



**Fig. 169: Schematic and TEM cross-section along the gate of (a) the FinFET devices and (b) stacked GAA-NW devices.**

The measurement results are depicted in Fig. 170. The nFET nanowire devices show ~54% increase in thermal resistance over the nFinFET devices.

For the current pGAA-NW devices, the S/D junction consists of *Si* instead of the typically used SiGe for pFETs. The fact that also in this case a higher sensor $\Delta T$ is measured in the pGAA-NW (a 33% increase w.r.t. their nGAA-NW counterpars), *which has no SiGe in the S/D but also an elevated measured series resistance on the source*, further confirms our theory about the electrical resistance delocalizing the heat source of the device.

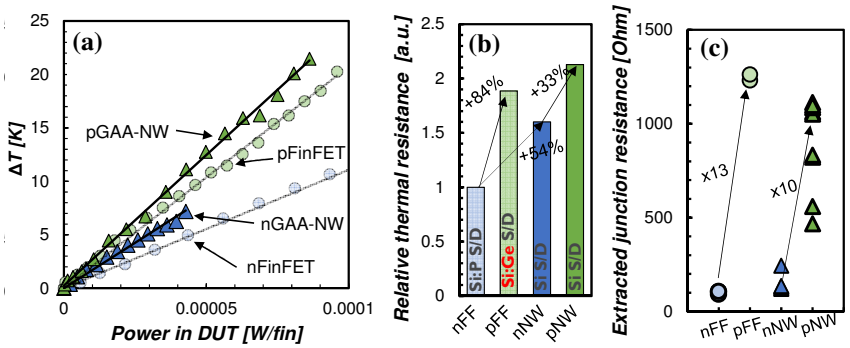



**Fig. 170. (a) Temperature increases versus power compared for FinFET and NW FETs. (b) Extracted thermal resistances for FinFET and NW of both polarities. (c) Extracted electrical resistance of the source contacts. Note the ~10-13x increased series resistance for the pFET devices.**

## 5.7 Further FinFET scaling

In this Section, we make further predictions for the self-heating effect in FinFETs in future technology nodes based on 3DFEM simulations.

The geometric dimensions for the 14nm node are according to available commercial technology parameters, whereas the dimensions for 10 and 7nm nodes are derived by scaling the contacted gate pitch and fin pitch by ~0.7x, while keeping gate length and fin width optimized for maintaining good electrostatics (*<70 mV/dec*). The fin length itself is considered long (~um order). All the utilized dimensions are depicted in Table VII.

**Table VII: Geometric technologic assumptions for bulk FinFET scaling from the 14nm node down to 7nm.**

| Technology node | 14nm | 10nm | 7nm |
|---|---|---|---|
| Contacted Gate Pitch [nm] | 90 | 64 | 42 |
| Gate Length [nm] | 30 | 24 | 18 |
| Fin Pitch [nm] | 48 | 36 | 24 |
| Fin Width [nm] | 10 | 7 | 5 |
| Fin Height [nm] | 30 | 30 | 35 |
| Target Ion per fin [uA/fin] | 70 | 70 | 70 |

### *5.7.1 Node scaling*

The results for the node scaling are shown in Fig. 171. Typical trends are observed: for smaller technology nodes, the thermal resistance will increase. On single fins, $R_{TH}$ increases by 20% per node since the device dimensions decrease, degrading the heat dissipation. As expected, $R_{TH}$ also increases with number of fins. For the fin height scaling, a constant current per fin was assumed. In this case, the taller fins are generated by increasing the fin height (i.e. in contrast to the previous Section where the fin height was increased by a deeper gate recess). Consequently, the simulation data is normalized to a constant current per fin.

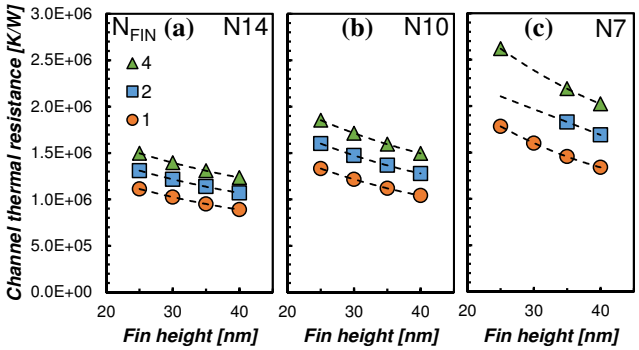



**Fig. 171: Simulated thermal resistance of the channel for (a) N14, (b) N10 and (c) N7 technology nodes, as a function of fin height (constant $I_{ON}$ per fin assumed) and for varying number of fins.**

Subsequently, the heat dissipation paths are compared. This is done by comparing the heat flux w.r.t. the total heat flux through relevant planes in the fin. For example: the heat flux towards the bulk is derived from the heat flux in the contact plane between the fin and the bulk.

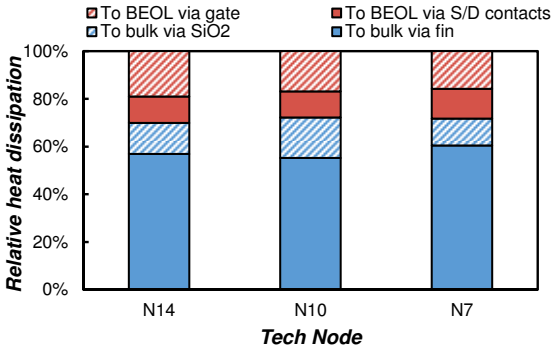



**Fig. 172: Relative contributions of heat dissipation for the various technology nodes remains more or less constant: ~70% of the heat dissipates towards the bulk.**

In the simulated FinFET devices, about 70% of the heat dissipates towards the substrate (via the fin and $SiO_2$) (Fig. 172). These *value remains rather*

*unchanged with scaling*. About 15% of the heat is dissipated via the contacts. In Fig. 173, we show the contributions of each of the thermal *branches*, from the channel towards the outside boundary conditions, redistributed by the materials. In other words, we breakdown each heat dissipation path into a series chain of $R_{TH}$'s.

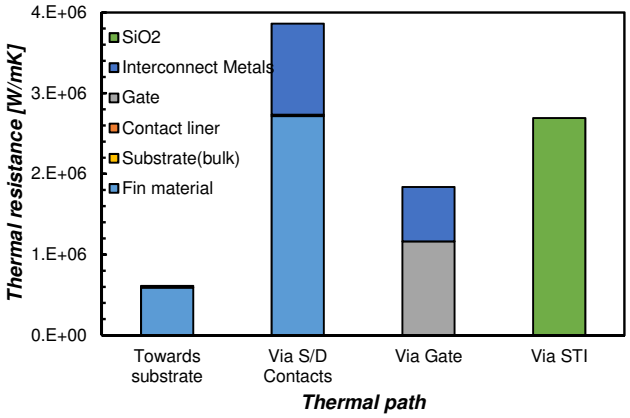



**Fig. 173: Example of the contributions of the thermal resistance in the fin towards the outside boundaries, calculated from the local ΔT and local flux.**

From this figure, we see that the branch towards the bulk has the lower overall thermal resistance, moreover, the impact of the substrate itself is neglible with respect to the impact of the Si fin towards the substrate. The second branch is the gate, in which the gate material has a strong contribution, but also the gate interconnect metal. Thirdly, the STI has a contributions which is larger than the S/D contacts. This is because the STI is ubiquitously present around the device, even though it has a low thermal conductivity.

Finally, and consistently with the previous results, the source/drain contacts form a high thermal resistance, making it a rather unlikely branch for heat dissipation. This is remarkable, because the S/D material is predominantly Si. This means, however, that there might be room for improvement: changing the gate-to-contact spacing could potentially reduce the thermal resistance of the device. This will be investigated in Section 5.7.2.

### 5.7.2 Gate-to-contact scaling and S/D recess

To simulate the impact of gate-to-contact scaling, the simulation environment has been improved to better represent the geometry of the contact, now being slightly recessed compared to the fin channel (Fig. 174). Also the S/D epi material has been replaced by SiGe.

We discriminate three cases: no recessed contact (i.e. the S/D EPI remains untouched, and the local interconnect is wrapped around the EPI), a half-recessed contact, (i.e. the local interconnect replaces half of the S/D EPI) and a full recess, (i.e. the local interconnect fully recessed the S/D EPI).

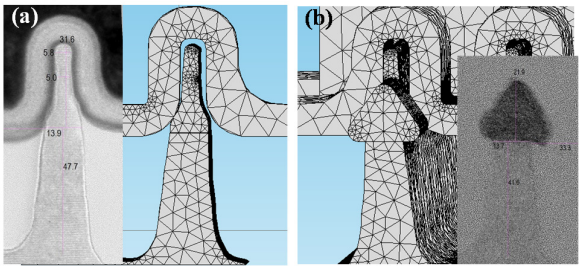



**Fig. 174: TEM micrographs and modeled fin geometry along the gate, of (a) the channel and the gate and (b) the SiGe contact region.**

By bringing the contact closer to the gate—even if it consists of SiGe—$R_{TH}$ is reduced (Fig. 175). The effect is stronger when the gate is fully recessed, which can be understood from the fact that the lowly thermally conductive SiGe is being replaced by local interconnect metal. In a similar way, we can also observe that increasing the recess depth has a stronger impact when the contact is brought closer to the gate. Finally, it can be observed that changing the recess depth has a slightly stronger effect in N7 than in N10.

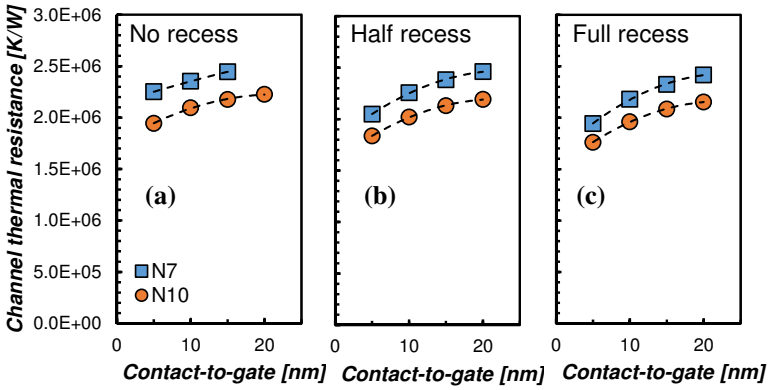



**Fig. 175: Simulated channel thermal resistance as a function of gate-to-contact distance and S/D recess depth for N10 and N7 technology nodes ($N_{FIN}$ = 2).**

# 5.8 Impact of SHE on performance and reliability

As we have discussed the self-heating effect in current nodes and made predictions for future nodes, an important aspect that remains to be discussed is the *impact* of the self-heating effect on the transistor drive current and reliability. In this Section, we will discuss how to model the impact of SHE on circuit performance and discuss several reliability aspects.

## *5.8.1 Impact on circuit performance*

The impact of the self-heating effect on the drive current will depend on the normal and reverse temperature dependence and concomitantly on the operating conditions of the device, as discussed earlier. Also, a major question to be solved is how the self-heating will have an *effect on GHz-frequency switching circuits*. An illustration of the variation of voltages and currents over time of FETs in a ring-oscillator circuit is given in Fig. 176.

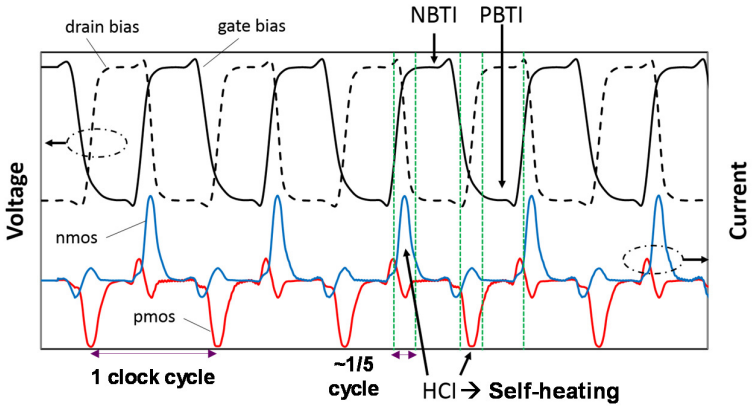



**Fig. 176: Variation of voltages and currents over time in a ring-oscillator circuit. The FETs have current cycles 5x faster than the operating frequency and a low duty cycle [courtesy of Pieter Weckx].**

The drive current of the device, taking into account the SHE, can be implemented in a BSIM4 model. This is done by implementing the temperature-induced $\Delta$ of the $I_D$ at the various $V_G$ and $V_D$ operating conditions (indicated as $\Delta(I_{DS}, V_{DS})$ in Fig. 177), which can be experimentally measured. The nFET shows a typical temperature trend where $I_{ON}$ *degrades* at high temperature (-5% at 125ºC) while pFET shows a reverse temperature dependence, with $I_{ON}$ *increasing* up to 7.5% at 125ºC (not shown here).

Subsequently, the device's thermal resistance $R_{TH}$ and its thermal conductance $C_{TH}$ are plugged into a model as parallel branches in a circuit that will mimic the FET's temperature over time, as illustrated in Fig. 177.

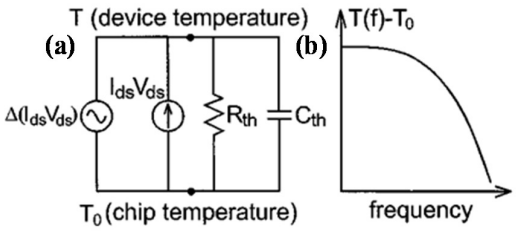



**Fig. 177: (a) RC network model to incorporate self-heating effects in a BSIM4 model (b) At high frequencies, the self-heating (indicated as *T*(*f*)) will become zero because the of thermal capacitance acting as a low-pass filter.**

Based upon this data, the transistors output conductance ($g_{ds}$) as a function of input frequency can be simulated. If RF-structures are available, the *$g_{ds}$-frequency* curve of the transistor can directly be measured, as depicted in Fig. 178 for a long channel (280nm) planar SOI device. The conductance difference at low and high frequencies can then be translated into a thermal impedance [Tenbroeck96]. Details on the extraction methodology can be found in [Rinaldi01]. In this example (Fig. 178), the model consists of a series network of RC-elements with a total $R_{TH} = 6.43e^3$ [K/W] and a $C_{TH} = 6.0640e^{-12}$ [J/K]

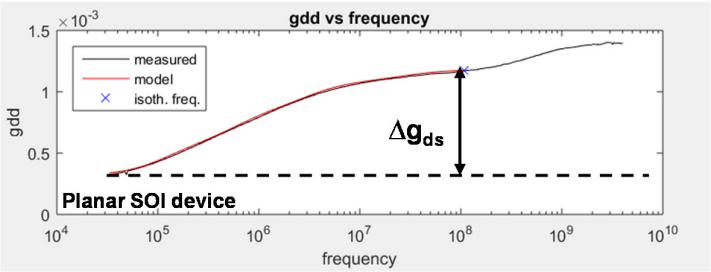



**Fig. 178: The conductance difference $\Delta g_{ds}$ at low and high frequencies can be translated into thermal impedance and modeled with RC-networks in series.**

In the example below, simulations for the delay of an inverter-based ring-oscillator are made, based upon the simulated thermal resistances from the Section 5.7. The results are used for various technology nodes (varying $N_{FIN}$,

fin height,…). The temperature dependence of $I_D(V_G, V_D)$ is calibrated from imec's 14nm FinFET technology, which shows a normal temperature dependence for nMOS and a reverse temperature dependence for the pFET devices at typical operating $V_{DD}$ of 0.8V.

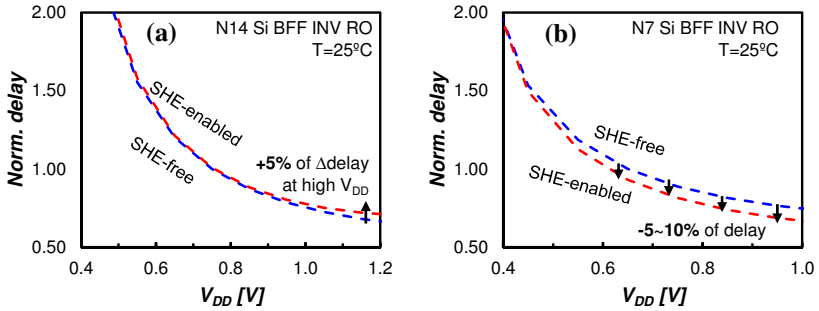



**Fig. 179: Normalized delay for a invertor-based ring-oscillator simulated for (a) N14 and (b) N7 parameters of thermal impedance [courtesy of Doyoung Jang]**

Fig. 179 shows the normalized delay for a invertor-based ring-oscillator circuit, simulated with and without self-heating effects with the BSIM4 model. From the results, we can observe that the impact of the self-heating effect depends on the operating conditions of the circuit. For N14, the self-heating will increase the delay by ~5%. In N7, the self-heating will decrease the delay over a wide range of operating conditions by 5-10% due to the strong reverse temperature dependence of the pFET. (Fig. 179(b)).

From these simulations, we can conclude that the effect of self-heating on circuit should be simulated case-by-case, and that even though thermal resistances of devices are increasing in scaled technology nodes, *the effect on circuit performance is not necessarily adverse*.

### *5.8.2 Impact on reliability*

Continuing the discussion from Chapter 2, it is known that many degradation effects are temperature activated, for example in BTI, the observed $\Delta V_{TH}$ was shown to follow Arrhenius' law. As discussed in Chapter 4, the temperature increases the stress induced-leakage current by enhanced

charge trapping and by enhanced trap-assisted-tunneling. Temperature is also known to enhance the recovery of degradation by increased charge detrapping [Aichinger13].

Most importantly, however, the above described degradation mechanisms are based on high gate bias stress and low drain bias, i.e. even with a high thermal resistance, the device *remains cold* (unless heated from a chuck) during these stress conditions.

The major exception from this is channel hot carrier stress. As discussed in Chapter 2, for short channel devices, it was shown that hot carrier degradation is temperature-activated, whereas discussed above, it was shown that narrower devices have a higher thermal resistance.

A possible experiment is to study the channel hot carrier degradation as a function of fin width. In Fig. 180, the $I_D$-$V_G$ after repeated stress cycles is shown. Whereas the 1000nm and 250nm wide devices can be considered planar-like, the 20nm device can be considered a real FinFET device.

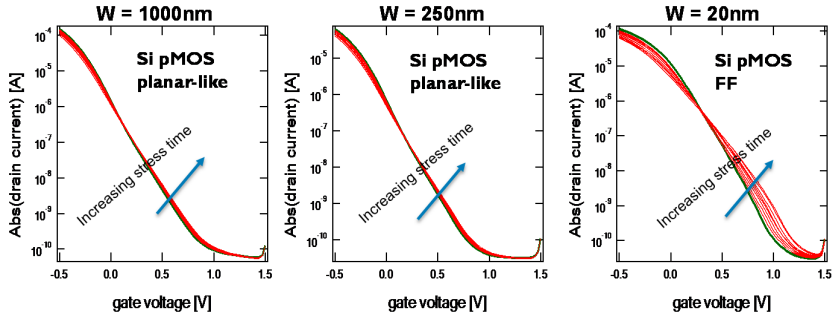



**Fig. 180: Additional CHC degradation is observed in narrow (FinFET) devices over wide (planar) devices at identical operating conditions.**

The interpretation of the experiment is however not straightforward. At the first sight, the higher degradation at elevated $V_{DD}$ could be attributed to self-heating effects in the FinFET device (Fig. 181). However, for the narrower devices, the ratio of the <110> (i.e. sidewall) versus <100> (i.e. top) surface is higher than for wide devices. Typically, both higher initial $D_{IT}$ as $\Delta D_{IT}$ are seen for devices with <110> sidewalls. The higher initial $D_{IT}$ is commonly attributed to the higher Si atom density on the <110> crystal orientation. The

increasing $\Delta D_{IT}$ is similarly ascribed to the higher availability of precursor defects [Cho09].

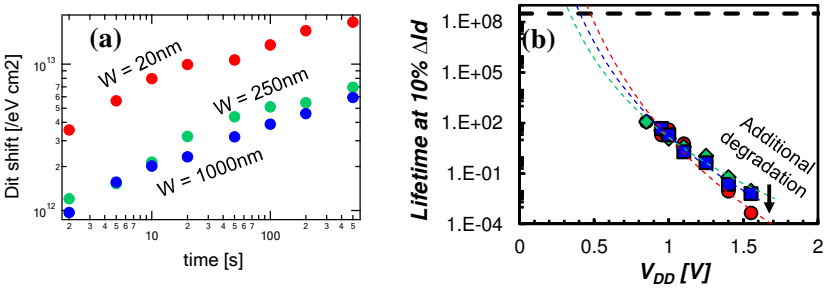



**Fig. 181: (a) Extracted $D_{IT}$ shows a higher initial $D_{IT}$ and $\Delta D_{IT}$ for FinFETs and (b) the lifetime extrapolation plot indicates identical lifetimes at lower $V_{DD}$ but reduced lifetime at elevated $V_{DD}$ due to additional degradation.**

From these kind of experiments, one cannot draw any conclusion about the impact of the self-heating effect on the degradation, because we cannot be certain that the devices have identical intrinsic properties, other than their thermal resistance. Also, the device's thermal resistance is typically overestimated since devices with large (i.e. hundreds) $N_{FIN}$ are utilized for CHC benchmarking. It are exactly those devices whereof we have shown that they exhibit a much larger $\Delta T$'s than low $N_{FIN}$ devices. Whereas it are the latter devices which typically used in circuit.

It remains clear though that the self-heating effect should be taken into account during CHC reliability testing. Empirical examples are given by Liu et al, proposing to compensate for the elevated temperature either physically, by lowering the environmental (i.e. chuck) temperature with increasing $V_{DD}$ conditions or mathematically, assuming Arrhenius' temperature activation [Liu14].

Our proposal to tackle this problem is to use arrays of devices in very high density, which allows parallel or separate stressing of the devices, capable of mimicking the low $N_{FIN}$ or the high $N_{FIN}$ devices. Moreover, such average would obtain statistics on individual CHC degradation of each device. For this purpose, we have designed an array consisting of over 5000 devices with tens

of shared drain lines, and with an addressable pass-gates for *each* device. Peripheral transistors on the drain lines have been avoided to remove series resistance effects. An example of the layout of the circuitry is given in Fig. 182.

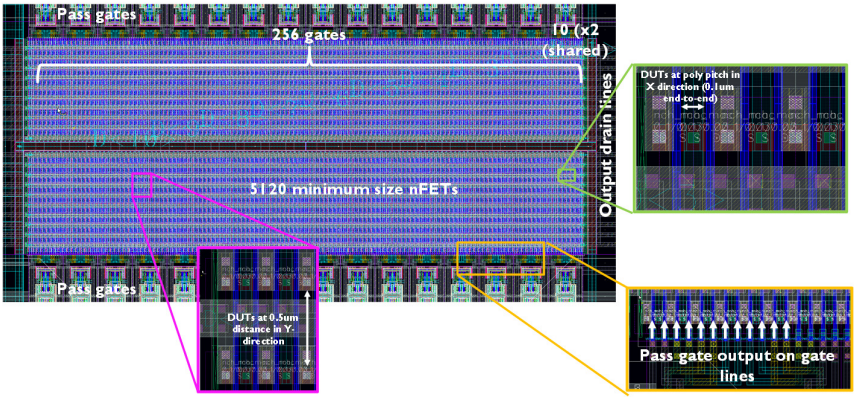



**Fig. 182: Layout of an array allowing separate or parallel CHC stressing of thousands of devices, needed to de-convolute self-heating effects from CHC degradation.**

## 5.9 Conclusions

In this Chapter, various measurement methodologies for measuring the self-heating effects were discussed and assessed for use in planar technology. We proposed a new methodology based on a heater-sensor configuration and corroborated those with finite-element simulations in planar devices. We showed that enhanced electro-thermal Mont-Carlo simulation techniques require proper boundary conditions to yield correct results. We showed that the heater-sensor structure allows to sense the drain-hotspot by reversing the bias conditions and corroborated this with multi-scale simulations.

The impact of architectural changes on the device thermal resistances was experimentally studied on imec's FinFET test vehicle. The EPI *composition* (SiGe in PMOS) causes major increase in self-heating effect. No clear impact of EPI implementation (raised vs embedded) on device self-heating effect was shown, but the recess depth of the local interconnect around the S/D junction

can have a significant impact on the $R_{TH}$. The measured thermal resistance of imec's first generation stacked GAA-NW is 60% elevated with respect to the baseline FinFET. We showed that increasing the fin height by a deeper gate recess reduced the thermal resistance.

Finally, we discussed a methodology to assess the impact of the self-heating effect on circuit performance by incorporating parameters in a BSIM4 model. We showed that the effect on circuit delay depends on the specific circuit and on the circuit's operating conditions, i.e. the supply voltage. However, at any of the investigated operation conditions for a typical ring-oscillator, the simulations indicate that the self-heating effect is currently not a limiting factor in FinFET transistors in terms of device performance.

Regarding device *reliability*, we showed that the impact of the self-heating effect on CHC is convoluted with other geometry-dependent effects, and thus cannot be straightforwardly extracted. It is however clear that more confined device geometries such as nanowires will give rise to worse SH which might cause enhancements of failure mechanisms such as stress-induced-leakage and oxide or junction breakdown. To separate the effects of the enhanced self-heating from additional degradation caused by *extrinsic* sources, we proposed and designed a high-density device array which will allow de-convolution of these effects.

# 5.10 References

[Aichinger13] Aichinger T. et al., "Application of On-Chip Device Heating for BTI Investigations", in "Bias Temperature Instability for Devices and Circuits", Springer, 2013.

[Aksamija13] Aksamija Z. and Knezevic I., « Thermal conductivity of Si1−xGex/Si1−yGey superlattices: Competition between interfacial and internal scattering", in Phys. Rev. B, Vol. 88, pp. 155318, (2013).

[Beppu12] Beppu N., Oda S., and Uchida K., "Experimental Study of Self-Heating Effect (SHE) in SOI MOSFETs: Accurate Understanding of Temperatures during AC Conductance Measurement, Proposals of 2ω Method and Modified Pulsed IV", IEEE International Electron Devices Meeting, pp. 642-645, (2012).

[Dallman95] Dallmann D. A. and Shenai K., "Scaling Constraints Imposed by Self-Heating in Sub-micron SO1 MOSFET's", IEEE Transactions on Electron Devices, Vol. 42, No. 3, pp. 489-496, (1995).

[Deng97] Deng F., Johnson R.A., Dubbelday W.B., Garcia G.A., Asbeck P.M. and Lau S.S., "Salicide Process for 400 A Fully-depleted SOI-MOSFETs using NiSi", in Proc. Of Int. SOI Conf., pp. 22-23, (1997).

[Enz95] Enz C., Krummenacher F., Vittoz E.A., "An Analytical MOS Transistor Model Valid in All Regions of Operation and Dedicated to Low-Voltage and Low-Current Applications, Analog Integrated Circuits and Signal Processing", in Journal on Low-Voltage and Low-Power Design. Vol 8, pp. 83-114, (1995).

[Fiegna08] Fiegna C., Yang Y., Sangiorgi E., and O'Neill A. G., "Analysis of Self-Heating Effects in Ultrathin-Body SOI MOSFETs by Device Simulation, IEEE Transactions on Electron Devices, Vol. 55, No. 1, pp. 233-244, (2008).

[Filanovsky01] Filanovsky I.M. and Allam A (2001) "Mutual compensation of mobility and threshold voltage temperature effects with applications in CMOS circuits", in IEEE Trans Circuits and Syst, Vol. 48, pp.876–884, (2001).

[Franco14] Franco J., Kaczer B. and Groeseneken G., "Poly-Si heaters for ultra-fast local temperature control of on-wafer test structures", Microelectronic Engineering, Vol. 114, pp. 47-51, (2014).

[Grayeli11] E. Bozorg-Grayeli et al.,"High temperature thermal properties of thin tantalum nitride films", Applied Physics Letters, Vol. 99, pp. 261906, (2011).

[Groeseneken10] Groeseneken G., Degraeve R., Kaczer B., Martens K., "Trends and perspectives for electrical characterization and reliability assessment in advanced CMOS technologies", IEEE ESSDERC proceedings, pp.64-72, (2010).

[Kawaski09] Kawasaki H., et al., "Challenges and solutions of FinFET integration in an SRAM cell and a logi circuit for 22 nm node and beyond," in Proc. Int. Electron Devices Meeting, pp.1–4, (2009).

[Kim11] Kim I., "Interface and Size Effects on TiN-based Nanostructured Thin Films", Doctoral dissertation,, Texas A&M University, (2011).

[Lee09] Lee J., Liao A., Pop E., and King W. P., "Electrical and Thermal Coupling to a Single-Wall Carbon Nanotube Device using an Electrothermal Nanoprobe", Nano Letters, Vol. 9, No. 4, pp. 1356-1361, (2009).

[Liu04] Liu W., and Asheghi M., "Phonon-boundary scattering in ultra-thin single crystal silicon layers", Applied Physics Letters, Vol. 84, pp.3819–3821, (2004).

[Liu06] Liu W., Etessam-Yazdani K., Hussin R. and Asheghi M., "Modeling and Data for Thermal Conductivity of Ultrathin Single-Crystal SOI Layers at High Temperature", in J. Heat Transfer, Vol. 128, No.1, (2006).

[Liu13] Liu J., Zhu J., Tian M., Gu X., Schmidt A., Yang R., "Simultaneous measurement of thermal conductivity and heat capacity of bulk and thin film materials using frequency-dependent transient thermoreflectance method", Review of Scientific Instruments 84, 034902, (2013).

[Lu09] Lü X., "Thermal conductivity modeling of copper and tungsten damascene structures", Journal of Applied Physics, Vol. 105, 094301, (2009).

[Lui14] Liu S.E. et al., "Self-Heating Effect in FinFETs and Its Impact on Devices Reliability Characterization", in Proc. IRPS, pp. 4A.4.1-4, (2014).

[Majumdar95] Majumdar A., Fushinobu K. and Hijikata K., "Effect of hate voltage on hot electron and hot phonon interaction and transport in a submicrometer transistor", in J. Appl. Phys., Vol. 77, pp. 6686-6694, (1995).

[Mertens16] Mertens H. et al., "Gate-all-around MOSFETs based on Vertically Stacked Horizontal Si Nanowires in a Replacement Gate Metal Gate Process on Bulk Si Substrates", VLSI Tech. Dig., pp. 158-159 (2016).

[Morshed09] Morshed T.H., et al., "BSIM4.6.4 MOSFET model user's manual". [Online] http://wwwdevice.eecs.berkeley.edu/~bsim3/bsim4_arch_ftp.html, (2009).

[Panzer09] Panzer M. et al., "Thermal Properties of Ultrathin Hafnium Oxide Gate Dielectric Films", Vol. 30, No. 12, (2009).

[Park95] Park C., et al.,"Reversal of temperature dependence of integrated circuits operating at very low voltages", Proc. IEDM, pp. 71–74, (1995).

[Prasad13] Prasad C. et al., "Self-heat reliability considerations on Intel's 22nm Tri-Gate technology", in Proc. IRPS, pp.5D.1.1-1.5 (2013).

[Qazi15] Qazi S.S. et al, "Multi-Scale Modeling of Self-Heating Effects in Silicon Nanoscale Devices", in Proc. IC on Nanotechnology, pp. 1461-1464, (2015).

[Quintana11] Alvarez-Quintana J., et al., "Thermal conductivity of thin single-crystalline germanium-on-insulator structures", in J. Heat and Mass Transfer, Vol. 54, pp.1559-1962, (2011).

[Raleva14] Raleva K., Bury E., Vasileska D., and Kaczer B., "Uncovering the Temperature of the Hotspot in Nanoscale Devices", accepted for 17th International Workshop on Computational Electronics, Paris, (2014).

[Rhyner13] Rhyner R., and Luisier M., "Self-heating effects in ultra-scaled Si nanowire transistors", IEEE International Electron Devices Meeting, pp. 790-793, (2013).

[Rinaldi01] Rinaldi N., "Small-signal operation of semiconductor devices including self-heating, with application to thermal characterization and instability analysis", in IEEE TED, Vol 48, No. 2, (2001).

[Scholten09] Scholten A., et al, "Experimental assessment of self-heating in SOI FinFETs", IEEE International Electron Devices Meeting, pp. 305-308, (2009).

[Stojanovic11] Stojanovic N., Berg J. M., Maithripala D. H. S. and Holtz M., "Direct measurement of thermal conductivity of aluminum nanowires," Applied Physics Letters, vol 95, pp. 091905-091905-3, (2011).

[Su94] Su L.T. et al., "Measurement and modeling of self-heating in SOI nMOSFET's", in IEEE Transactions on Electron Devices, Vol 41, No 1, pp. 69-75, (1994).

[Synopsys14] Sentaurus TCAD – Synopsys user manual, available online. (2014).

[Sze81] Sze S. M., "Physics of semiconductor devices", 2nd ed. John Wiley and Sons, (1981).

[Takahashi13] Takahashi T., Matsuki T., Shinada T., Inoue Y. and Uchida K., "Comparison of Self-Heating Effect (SHE) in Short-Channel Bulk and Ultra-Thin BOX SOI MOSFETs: Impacts of Doped Well, Ambient Temperature, and SOI/BOX Thicknesses on SHE", IEEE International Electron Devices Meeting, pp. 184-187, (2013).

[Tenbroeck96] M. Tenbroeck, et al, "Self-Heating Effects in SOI MOSFET's and Their Measurement by Small Signal Conductance techniques", IEEE Transactions on Electron Devices, Vol. 43, No. 12, pp. 2240-2248, (1996).

[VanOverstraeten73] Van Overstraeten R.J., Declerck G. and Broux G.L. "Inadequacy of the classical theory of the MOS transistor operating in weak inversion" in IEEE Trans. Electron Dev., Vol. 20, pp. 1150, (1973).

[Vasileska12] Vasileska D., Raleva K., Hossain A., and Goodnick S.M., "Current progress in modeling self-heating effects in FD SOI devices and nanowire transistors", J Comput Electron, Vol. 11, pp. 238-248, (2012).

[Wolpert08] Wolpert D. and Ampadu P, "Normal and reverse temperature dependence in variation tolerant nanoscale systems with high-k dielectrics and metal gates", in ACM Int Conf on Nano-networks, pp. 1–5, (2008).

# Chapter 6: Self-heating considerations for future technology nodes

In this Chapter, we show that the implementation of various alloys such as SiGe or GaAs to mechanically stress or act as a strain-relaxed buffer for the channel, can have a dramatic impact on the device self-heating effect (SHE), both in FinFET as in nanowire devices.

## 6.1 Introduction

Future technology nodes will become more complicated as the silicon channel may be replaced by high-mobility materials like Ge/SiGe or III/V-semiconductors. For example, the potential of strained Ge p-FinFETs on a SiGe Strain Relaxed Buffer (SRB) has been demonstrated in recent years using either STI-last [Mitard14] or replacement channel [Witters13] integration options, each technique yielding various thicknesses of the SiGe SRB. The advantage of the latter approach is, however, the easier co-integration with other channel materials, for instance III/V channels.

III/V materials, on the other hand, have attracted much attention as a potential high mobility channel material for advanced scaling nodes in n-FinFETs. Significant progress has been made to integrate III/V on a VLSI compatible platform with different techniques. Unlike the case of Ge, which can be integrated in a slightly modified semiconductor processing line, using similar know-how as for Si, this is not the case for processing III/V compounds. Recently however, InGaAs GAA-NW processed on large scale Si wafers [Waldron15], [Zhou16], with effective gate lengths of 36nm have been demonstrated. In parallel, other channel materials (e.g. InGaAs vs InAs) and device structures (e.g. FinFET or GAA-NW) are still being considered, depending on their optimum performance, which is typically also strongly dependent on external factors. An example of such an external factor is the surface passivation, which has a strong impact on their performance and also on the charge trapping in border traps [Franco16]. Fig. 183 shows schematics

of FinFET and GAA-NW device architectures with various new semiconductor materials.

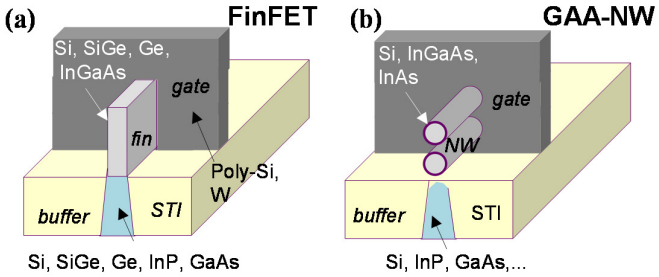



**Fig. 183: For future (a) FinFET or (b) GAA-NW nodes, various new semiconductor materials will be introduced for their direct (i.e. as channel material) or indirect mobility-enhancing (i.e. by strain) properties.**

In this Chapter, we first discuss the thermal properties of semiconductor compounds and alloys. Subsequently, we describe the experimental setup and measurement methodology for gate resistance thermometry. Finally, we make three case studies of self-heating in future devices: Ge FinFETs with SiGe strain relaxed-buffers, InGaAs-based FinFETs and we conclude with two generations of InGaAs-based GAA-NW.

## 6.2 Thermal properties of compounds and alloys

The total thermal conductivity in semiconductors consists of a *lattice* and an *electronic* contribution. The lattice thermal conductivity is the dominant mechanism over a wide range of carrier concentrations in silicon, germanium, and several III-V semiconductors, and is commonly modeled as a temperature dependent power law [Palankovski00]. Only in heavily doped semiconductors, the electrons or holes can contribute to the thermal conductivity. The semiconductor's thermal conductivity $\kappa(T)$ can then be expressed as:

$$\kappa = \kappa_{300}\left(\frac{T}{T_0}\right)^{\alpha_\kappa} \tag{6.1}$$

with $\kappa_{300}$ the thermal conductivity at room temperature, and $\alpha^\kappa$ the power-law exponent.

In the case of alloys (e.g. $Si_{1-x}Ge_x$), $\kappa$ will vary as the composition changes between the basic materials (i.e. Si and Ge in this case). A harmonic mean is used to model $\kappa_{300}$, with an additional bowing factor $C_\kappa$ to account for the drastic reduction of the thermal conductivity for the alloy [Palankovski00]:

$$\kappa^{A_{1-x}B_x} = \frac{1}{\frac{1-x}{\kappa^A{}_{300}} + \frac{x}{\kappa^B{}_{300}} + \frac{(1-x).x}{C_\kappa}} \tag{6.2}$$

The temperature dependence for alloys can be derived in a similar way as in Eq. 6.1, replacing $\kappa_{300}$ with $\kappa^{A_{1-x}B_x}$, with the power-law exponent the $\alpha_\kappa$ typically linearly interpolated between both materials' parameters.

Thermal conductivities for a few typically used high-mobility compounds and alloys are depicted in Fig. 184 (a), and their material composition dependence is depicted in Fig. 184 (b).

From this picture, it is clear that in semiconductor *alloys* such as SiGe or InGaAs, there is a drastic reduction in thermal conductivity with respect to their constituent materials, whereas in *compounds*, such as InP or GaAs, this is not the case. Alloys are typical mixes of elements, resulting in a solid solution. A compound however is an association of several elements bound together by chemical bonds. It is not possible to obtain a compound by just alloying, but they are only achievable through specific chemical reactions.

The main explanation of this difference in thermal conductivities is given in Fig. 185, which shows the crystal configuration for a compound and an alloy: in compounds, phonons can travel and carry thermal energy mostly unperturbed throughout certain planes in the lattice, whereas in alloys *the complete random mix of the constituent atoms* are causing the phonons to encounter randomly varying masses as they travel through the structure. It is this *aperiodic mass variation* between the two constituent types of atoms

which perturbs the lattice waves and leads to strong alloy scattering of phonons, therefore also called *mass disorder scattering*.

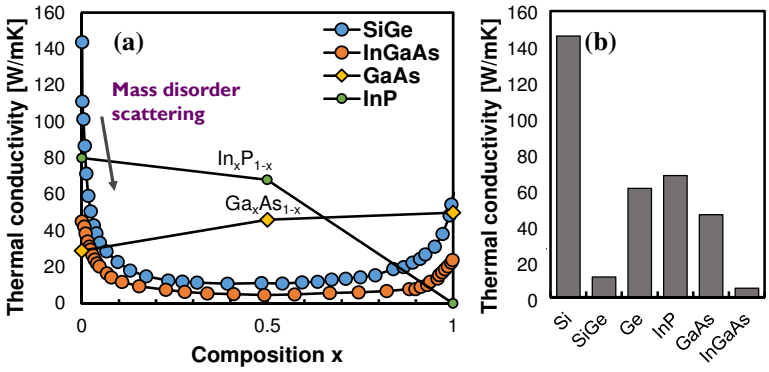



**Fig. 184: (a) Bulk thermal conductivity values for various high-mobility semiconductors and (b) the strong difference in thermal conductivity between semiconductor *compounds (GaAs, InP)* and *alloys (SiGe, InGaAs)* [replotted from Palankovski00].**

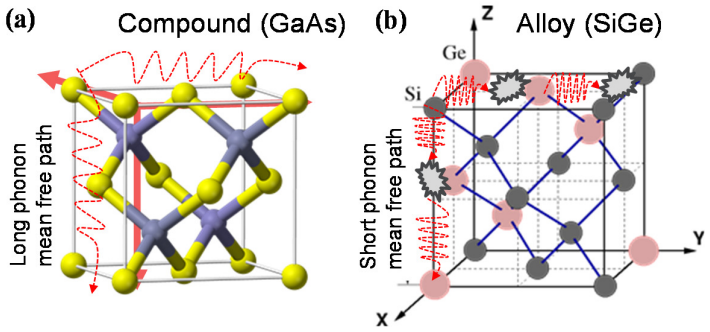



**Fig. 185: (a) In compounds, phonons can travel and carry thermal energy mostly unperturbed whereas in (b) alloys the random mix of the constituent atoms causing the phonons to encounter randomly varying masses as they travel through the structure.**

The physical mechanism responsible for heat transfer—phonon transport and interactions—will thus depend on material composition, temperature,

surface orientation and proximity, and doping. In this case, the thermal conductivity is derived from summing over all the phonon momenta and branches, which can in turn be obtained by solving the time-independent phonon Boltzmann Transport Equation (PBTE). Details of this method are given by [Aksjamija13]. This will result in direction-dependent thermal conductivity, illustrated in Fig. 186.

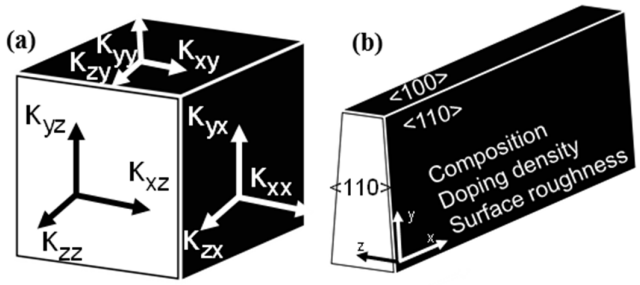



**Fig. 186: Thermal *tensors* are generated "ab-initio", based on (b) composition, geometry, surface roughness and doping concentration of the fin. The tensors generated for the fins are an order of magnitude smaller than the bulk conductivities, while the *alloy* conductivities are again an order of magnitude smaller than the pure materials.**

The anisotropy is caused by the difference in the effect of interface scattering on in-plane versus cross-plane transport: in the in-plane direction, heat is carried largely by phonons whose velocities are nearly parallel to the interface and which therefore undergo significantly less interface scattering than phonons that carry heat cross-plane. The anisotropic nature of this phenomenon is implemented in using the mathematical construct of geometry-specific thermal conductivity *tensors*.

As an example, thermal conductivity tensors are generated for bulk Si and Ge, but also the SiGe alloy (Fig. 187). Also the resulting tensors for thin-film (20nm) fin with <110> sidewalls are calculated. The thermal resistance is strongly reduced for SiGe and reduced even more for thin film material. The temperature dependence of thermal conductivity tensors can then be self-consistently included in 3DFEM simulations, albeit that this dependence is very low for SiGe FinFETs due to the additive nature of the above described scattering phenomena.

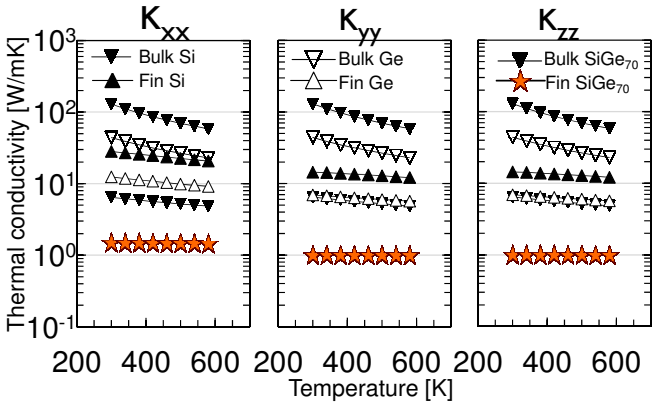



**Fig. 187: The resulting thermal conductivities for bulk material and a 20nm fin in the various directions. The results are in line with for above results are in line with other references such as [Cheaito12].**

## 6.3 Experimental description and measurement methodology

In this section, we will describe the experimental setup, i.e. the designed devices and the gate resistance thermometry methodology.

### *6.3.1 Design of experiment*

For gate resistance thermometry, the gate resistances are measured using a 2-point connection at both ends of the gate [Mautry90]. We have designed devices in various geometries, which are tabulated in Table VIII. The number of fins/nanowires are varied and also the fin/wire width (Fig. 188).

It should be noted that one important parameter is the ratio of the total gate poly-length (as indicated in Fig. 188 (b)) with respect to the active gate (i.e. the part of the gate that is located on top of an active fin). This ratio plays an important role in the sensitivity of the measurement. Due to lithographic constraints, this ratio cannot be kept constant and is much larger for single fin

devices than for their 10-fin equivalents. This signifies that a smaller fraction of the gate will be heated in the single fin device. The obtained measurement results for self-heating in single fin versus 10-fin devices can thus only be compared by additional simulations.

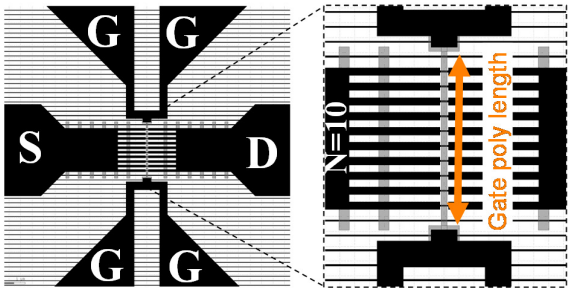



**Fig. 188: Layout of a FinFET ($N_{FIN}$=10) with a gate with 4 connections.**

**Table VIII: Combinations of test structure parameters studied.**

| $W_{FIN}$ [nm] | Fin pitch* [nm] | Gate Poly length [um] | | |
|---|---|---|---|---|
| | | $N_{FIN}$ = 1 | $N_{FIN}$ = 4 | $N_{FIN}$ = 10 |
| 20 | 200 | 0.65 | 1.25 | 2.45 |
| 30 | 210 | 0.7 | 1.33 | 2.58 |
| 40 | 250 | 0.9 | 1.65 | 3.15 |
| 75 | 460 | 2 | 3.31 | 6.09 |

*For $N_{FIN}$ > 1

## 6.3.2 *Measurement setup*

To measure the gate resistance, a fixed small differential voltage $\Delta V_{H\text{-}L}$ on top of the gate bias $V_G$ is applied. As a result, a current $I_{RES}$ will flow from one side to the gate to the other side. The resulting biases on the contacts will therefore be $V_G \pm \Delta V_{H\text{-}L}/2$, whereas the average gate bias remains $V_G$. The gate resistance can then be measured on-the-fly. However, for experimental FinFETs and corresponding gate stacks, poor interfacial layer and bulk oxide quality could lead to enhanced gate leakage by trap-assisted tunneling, as

described in Chapter 4. In that case, at elevated $V_G$, additional gate leakage current can jeopardize the measurement. To prevent this effect, the $\Delta V_{H\text{-}L}$ is swept for each $V_G$ bias condition, and the gate resistance is extrapolated from the slope of the $I_{RES}$-$V_{H\text{-}L}$ curve.

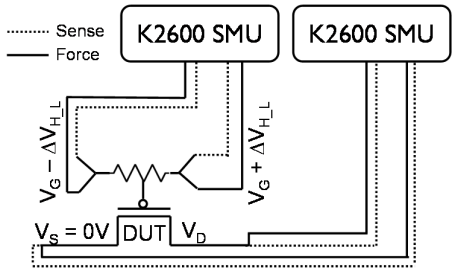



**Fig. 189: Schematic of measurement setup for gate resistance thermometry.**

The measurement schematic is depicted in Fig. 189. The measurement of the gate resistance is performed with a single 2-channel Keithley SMU 26xx unit, utilizing its intrinsic force and sense compensation capability on the gate contacts. A second linked 26xx supplies the source and the drain biases to dissipate power in the device.

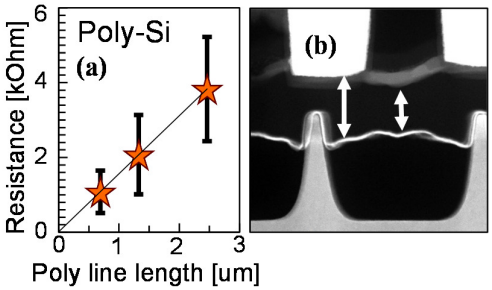



**Fig. 190: (a) The extracted electrical resistance as a function of poly line length. The median gate resistance scales excellently with poly-Si length. (b) The large device-to-device variation originates from poly-Si line roughness and granularity.**

The offset gate leakage and the characteristic resistance are measured for each device. Fig. 190 shows the extracted gate-resistance for Poly-Si gates as

a function of gate length. Even though the median gate resistance scales excellently with poly line length, a large device to device variation is observed. We attribute this variation to which we attribute to gate roughness and granularity. In Fig. 191 we show the principle for extraction of the TCR, utilizing the thermo-chuck. It can be observed that the devices show a first-order dependence of their resistance with temperature.

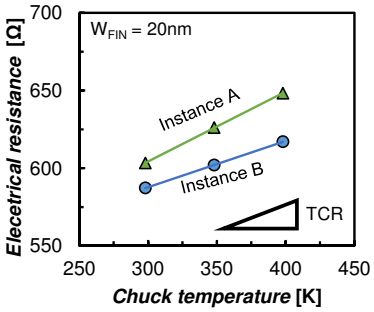



**Fig. 191: Principle of extraction of the TCR utilizing a thermo-chuck on a Poly-Si gate.**

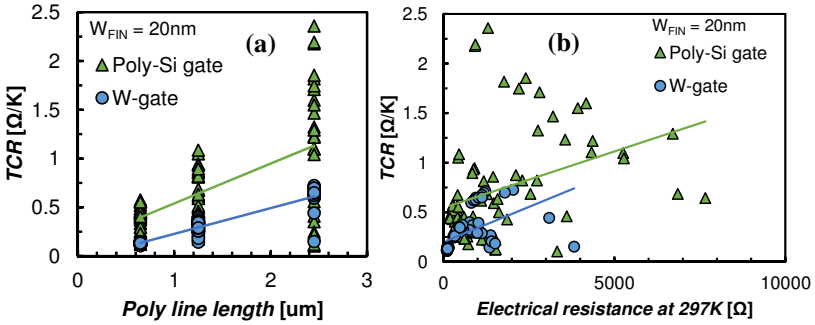



**Fig. 192: (a) TCR extracted as a function of poly line length for Poly-Si and W-gates and (b) TCR as a function of the electrical resistance at room temperature.**

In Fig. 192, we show the extracted temperature coefficient of resistance (TCR) for both Poly-Si and W gates. The poly-Si gate shows the strongest temperature dependence but also the largest variation. Also the initial electrical resistance in (b) cannot be used to predict the TCR of that particular

gate. Therefore, these variations necessitate an individual TCR calibration using the thermo-chuck for every device.

In a final stage, the $\Delta R$ is measured during operation of the transistor for various $V_G$ and $V_D$ conditions. This $\Delta R$ is converted to temperature increase in the self-heated transistor ($\Delta T$) utilizing the above-obtained TCR for that particular device. Fig. 193 shows the $\Delta T$ w.r.t. room temperature due to self-heating for every $V_G/V_D$ combination, with maximum voltages exceeding operating values to increase resolution. From this picture, we can observe that the largest $\Delta T$ appear during CHC stress conditions (up to 80K), whereas the device remains cold during BTI stress conditions.

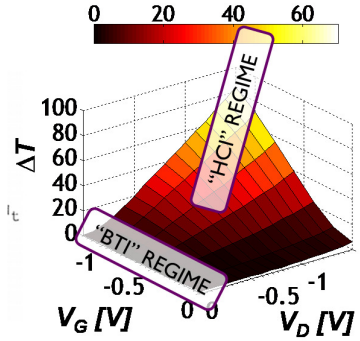



**Fig. 193: Measured values for various $V_D$ and $V_G$ conditions, showing the largest $\Delta T$ appearing during CHC stress conditions.**

The linear temperature dependence of the resistance allows direct extraction of the average gate $\Delta T$, which we confirmed using a coupled 3DFEM simulation. Firstly, we take into account a linear temperature dependence of the gate electrical *resistivity* $\rho$. This was confirmed with our measurements in Fig. 191 as the *resistance* increases linearly with a quasi-uniform temperature, provided from the thermos-chuck. Note that in this case resistance and resistivity are equivalent because the geometry of the gate is not changing in the experiment.

The 3DFEM simulation then couples both the electrical effects (electrical resistance and resistivity of the gate) as the thermal effects (Joule heating in the gate). In the simulator, we apply an electrical bias on the gate and we

extract the overall gate resistance based upon the simulated voltage drop along the gate. We simulated the following cases:

- During the calibration the heat profile is quasi-uniform and equal to the temperature of the substrate. The minor non-uniformity is caused by Joule heating because of the current flowing through the gate for the measurement itself.
- During the self-heating measurement, a strong non-uniform temperature profile is expected due to the localized heat generation in the fins, whereas the substrate itself remains cold.

The result of this measurement is shown in Fig. 194, which shows the heat profile along the gate during the calibration with the thermal chuck (i.e. quasi-uniform) or during transistor self-heating (i.e. with a strong gradient). From (c), we can observe that, regardless of the temperature *profile* in the gate, the measured *resistivity only depends on the average temperature of the gate,* and is shown not to be impacted by the temperature *gradient*.

This conclusion finalizes our experimental methodology for use as benchmarking tool further along in this Chapter.

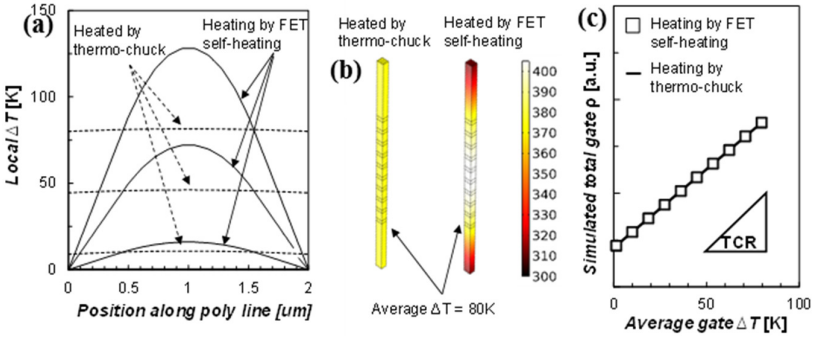



**Fig. 194: (a) A simplified electro-thermal simulation showing the heat profile along the gate during the calibration with the thermal chuck (i.e. quasi-uniform) or during transistor self-heating (i.e. with a strong gradient). (b) Illustration of the heat distribution along the gate from the 3DFEM simulations. (c) The total gate resitivity $\rho$ depends on the *average* temperature of the gate but is independent of the temperature gradient along that gate.**

## 6.4 Ge-channel FinFETs: impact of SiGe SRB

SiGe and Ge are foreseen as enablers of high-mobility channels in future FinFET, because of significant advantages: SiGe and Ge offer increased hole mobility required for next-generation CMOS technology, and they can be integrated in CMOS with a single high-k/MG stack [Mitard14]. Moreover, stress simulations predict that strain-relaxed-buffer layers are especially effective as strain booster compared to classical source/drain (S/D) stressors [Eneman12], thereby giving rise to additional mobility enhancement. Finally, Ge and SiGe channel devices are shown to exhibit excellent BTI reliability [Franco10].

### *6.4.1 Device description*

Fig. 195 shows TEM micrographs and schematic representations of the Ge-channel FinFETs studied here. The first device type, the fin consists entirely of Ge, i.e. there is no additional strain booster on the channel, and we refer to it as 'relaxed Ge'. The second device type consists of a SiGe strain-relaxed buffer (SRB) to stress the Ge channel (sGe/SiGe). The SRB is located in bottom-side of the fin. Details of the device fabrication can be found in [Bury15]. Finally, we have an equivalent full-Si FinFET device which we will use as a reference (no TEM image shown) for this case study.

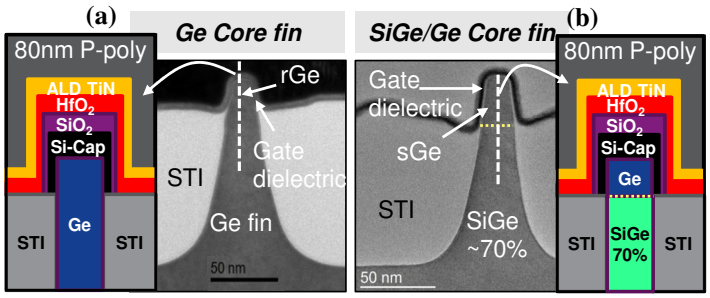



**Fig. 195: TEM micrographs and schematic representations of the high-mobility FinFET stacks studied. (a) In the first device type, the fin consists entirely of relaxed Ge (rGe). (b) The second type consists partially of a SiGe strain-relaxed buffer (SRB) to stress the Ge channel (sGe/SiGe).**

### 6.4.2 Measurement and simulation results

An illustration of *simulation* results of devices with $N_{FIN}$ = 10 and $W_{FIN}$ = 20 is shown in Fig. 196. Qualitatively, it is clear that the sGe/SiGe FinFET heats up much more than its Ge and Si counterpart. Also the strong temperature gradient on the outer parts of the gate is obvious. From these simulations, it is also clear that the gate is a good sensor, i.e. the sensed temperature comes close to the maximum temperature obtained device. Along the fins and the interconnect metals a strongly decaying temperature is seen.

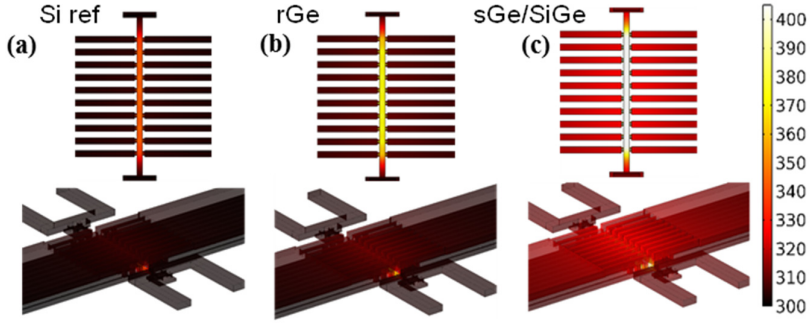



**Fig. 196: Top and 3D-views of the temperature profiles in $N_{FIN}$=10 devices for the (a) Si ref, (b) rGe and (c) sGe/SiGe channel FinFET. The sGe/SiGe channel device clearly heats up more than the rGe and Si ref.**

The results of the *measurements* for $N_{FIN}$=10 devices with the various material configurations, now also including $W_{FIN}$ variations, are shown in Fig. 197. Projected on top of the measurement data is the extracted gate temperature from the 3DFEM simulations. A good agreement is shown with the experimental data. Both measurements as simulations shows a linear dependence of $\Delta T$ versus power dissipated in the device in the entire measurement range, because the low temperature dependence of the thin-film thermal conductivities are reflected in a negligible non-linearity of the simulated thermal resistances and is not observable in the measurements or the simulations.

Fig. 198 shows the exhaustive overview of all the extracted $R_{TH}$'s for all the available $W_{FIN}$, $N_{FIN}$ and pitch combinations together with the simulation data. The adverse effect on the $\Delta T$ of decreasing $W_{FIN}$ and increasing the $N_{FIN}$

is shown for material combinations. It should be noted that he simulation results are obtained with *one single set* of thermal conductivity tensors for bulk and thin-film Si, SiGe and Ge , as introduced in Section 6.2.

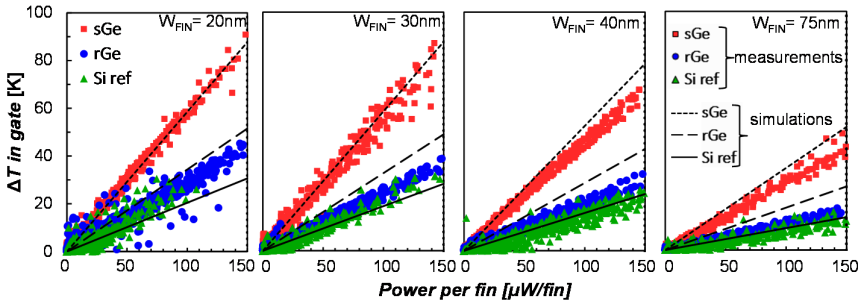



**Fig. 197: Measured and simulated gate $\Delta T$ with a single set of "ab-initio" tensors for varying fin widths show a linear thermal behavior over the entire power range and good qualitative match.**

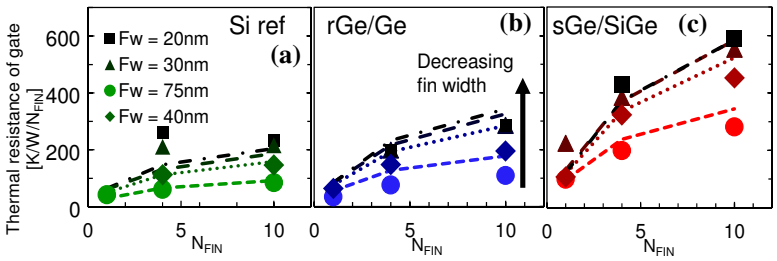



**Fig. 198: The simulations (lines) performed with a single set of "ab-initio" tensors match well for the (a) Si ref, (b) rGe and (c) sGe/SiGe measurements (markers), including the all device pitch and fin number combinations.**

## 6.4.3 Discussion

From the above measurement, it is clear that particularly the SiGe alloy in the bottom part of the sGe fin is impeding heat transfer towards the bulk thereby increasing the thermal resistance of the device.

In this discussion, we assume that the thermal behavior of the device is roughly equivalent to a simple resistive network, as depicted in Fig. 199(a). The temperature in the device corresponds to the potential on that node, whereas the heat flux can be represented as a current. The respective potential drops along the various resistances represent the drops in the temperature. From our earlier simulations in Chapter 5, we know that the main contributor for the self-heating effect is the thermal resistance created by the fin from the channel towards the bulk, indicated as $R_{TH\_FIN}$ in the illustration. Apart from that, another heat dissipation path is along the gate, from which we can only estimate its proportion via simulations. We can however assume that regardless of the fin material, this thermal resistance towards the gate will remain constant and only the $R_{TH\_FIN}$ will change for the various material options.

Consequently, the ratio of $\Delta T_{CHANNEL}/\Delta T_{GATE}$ is proportional to the amount of heat that can escape towards the bulk of the device. From the outcome of the simulations above, we can extract this ratio in Fig. 199. We can observe that for the higher number of fins, the ratio between the actual device temperature and the gate temperature is decreasing, thus making the gate a better sensor. Moreover, if we compare the cases for $W_{FIN}$ = 75nm, we can see that this ratio is reducing from ~7 for pure Si FF towards ~5 for sGe/SiGe FF. In the case for the narrow fins, with $W_{FIN}$ = 20nm, the ratio ranges from ~6 to 5.5 respectively. In other words the difference of thermal conductance towards the bulk over the various process flows is *mitigated as the fin pitch is decreasing*.

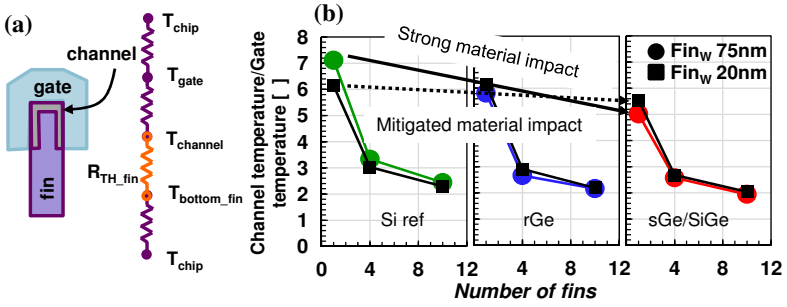



**Fig. 199: (a) The channel $\Delta T$ relative to the gate $\Delta T$ quantifies the impact of the bottom/buffer part of the fin as thermal barrier. (b) For narrower fins, the relative impact of the low thermally conducting SiGe buffer is mitigated.**

The main origin for this is the increased geometrical confinement, thereby reducing the relative importance of the fin as major thermal heat sink towards the bulk.

Note that another consequence of implementing SiGe in the fin is that more heat will escape towards the back-end-of-line (via the gate and/or the S/D contacts). This ranges the concern for BEOL reliability issues such as electromigration at the lowest interconnect levels, as this mechanism is typically strongly temperature activated [Cheng02].

### *6.4.4 Projections for future nodes*

In Fig. 200, the acquired channel $R_{TH}$'s are converted by new 3DFEM simulations taking into account the geometry according to the N7 node specifications discussed in Section 5.7.1. Also the VDD operating conditions and projected drive currents are taken into account.

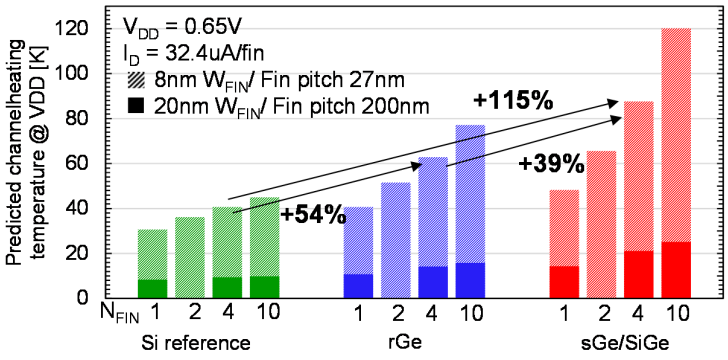



**Fig. 200: The impact of rGe and sGe/SiGe process flows on channel ΔT in the relaxed-pitch FinFETs extracted here (solid bars) and projected-size N7 FinFETs (dashed bars) at corresponding operating conditions.**

Consistent with our earlier observations, it can be observed that e.g. for a 4-fin device, this predicted temperature increase in rGe is 54% higher than the silicon reference fin, while the sGe with SiGe buffer is 115% increased, in contrast with the measurement data on more relaxed FF where a 350% increase was observed. The effect of utilizing low thermally-conductive alloys is thus still strongly present, but mitigated more scaled device nodes.

### *6.4.5 Conclusions*

Simulations and measurements have shown that high-mobility materials and particularly alloys, such as a SiGe buffer for strain-induced mobility enhancement in Ge, bring along a severe penalty on the thermal properties of 3D devices. N7 sGe pMOS FinFETs on SRB buffer of SiGe are shown to have 54% higher self-heating compared to rGe channels, and 115% compared to a Si reference FinFET.

## 6.5 III-V FinFETs: impact of buffer materials

In order for the promise of III/V to be realized as a viable option for CMOS co-integration with SiGe or Ge channel devices, these devices and/or materials must be integrated on 300mm or larger Si substrates in a fully VLSI compatible flow. The devices studied in this section are fully functional

InGaAs channel devices fabricated on 300mm Si substrates using the replacement fin technique. Part of this technique is that a SRB layer is used to grade up from Si towards the lattice constant of the III/V channel, and keep potential lattice mismatch defects away from the active channel. In this section, we will study the effect of the buffer material on self-heating in III/V devices.

### 6.5.1 Device description

Creating these InGaAs channel devices is a considerable challenge. Typically, a lattice mismatch of 8% between InP and InGaAs on one side, and the Si substrate on the other hand is expected. In a replacement fin process, the lattice misalignment defects can be trapped at the STI sidewall [Fitzgerald91]. Two devices proposals including TEM data are illustrated in Fig. 201. One device has a GaAs SRB, whereas the other device has a InP SRB. Other details on the device fabrication can be found in [Waldron14].

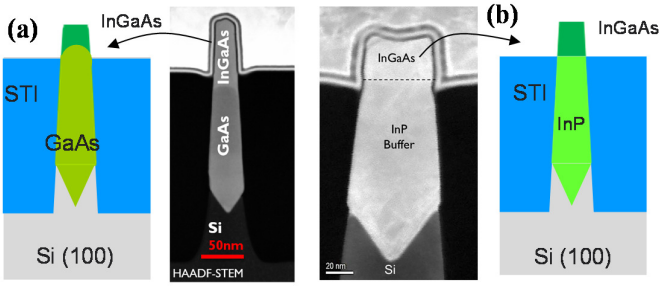



**Fig. 201: TEM micrographs and schematic representations of the high-mobility FinFET stacks studied. Both FinFETs exhibit InGaAs channels with either (a) a GaAs or (b) an InP SRB.**

From a thermal perspective, it is known that InP has slightly better thermal properties than Ge, and 47% higher than GaAs [Palankovski00]. We can thus expect that the InGaAs/InP devices will exhibit better thermal properties than their InGaAs/GaAs buffer counterparts. Finally, the InGaAs used in channel exhibits the typical very low thermal conductivity expected for non-dilute alloys, as was described in Section 6.2.

### 6.5.2 Measurement results

The yield of the wide-fin devices is very low. Therefore, only a subsection of narrow fin devices are utilized in this study, and we compare the data with the data obtained on Si/SiGe/Ge pFETs in Section 6.4.

The $\Delta T$-$P$ data (Fig. 202) of the InGaAs/InP FET show a non-linearity. It appears that a parasitic device is enabled at high $V_{D,}$ which can be observed from the unexpected increase in the $I_D$-$V_D$ at elevated $V_D$. Separating the $\Delta T$-$P$ data before and after the unnatural increase in $I_D$-$V_D$ roughly corresponds to two observed slopes. This could be an indication of a parasitic channel, which exhibits a lower thermal resistance since it is further removed from the gate and located closer to the bulk. To extract the thermal resistance of the InP SRB FinFET, the contributions of the real and the parasitic channel should be carefully de-convoluted.

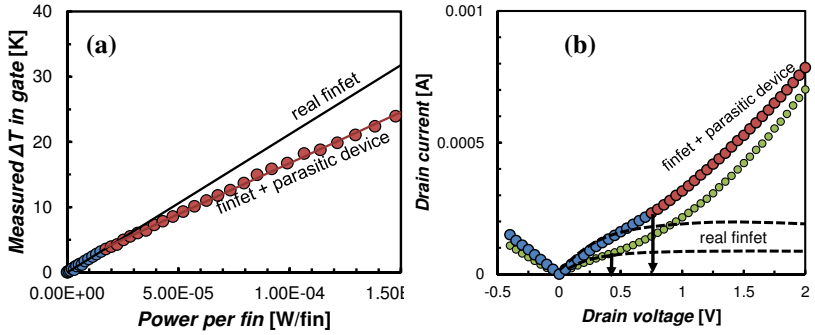



**Fig. 202: (a) the thermal resistance $R_{TH}$ for an InGaAs/InP FF shows non-linear behavior. (b) The $I_D$-$V_D$ characteristic of the device shows superlineair increase of current at elevated $V_D$, for both $V_G$ conditions. The difference between both curves in the non-gate-modulated region remains approximately constant.**

Under the assumption that leakage current of the parasitic channel is not modulated by the gate (which can be fairly made given the fact the difference between both curves in the non-gate-modulated region remains approximately constant), the $R_{TH}$ can be exactly extracted by utilizing the $\Delta I_D(\Delta V_G)$, which

only contains 'gate modulated' current ($\Delta I_D$), i.e. this current which is effectively flowing in the fin. The principle is illustrated in Fig. 203.

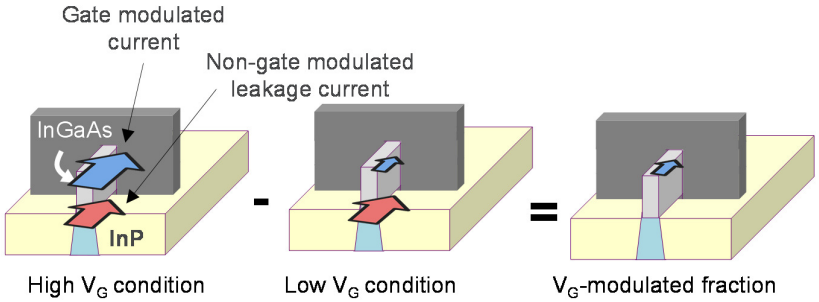



**Fig. 203: Principle of extracting only the gate-modulated current ($\Delta I_D$) by probing the device at various $V_G$ conditions.**

Therefore, the $\delta\Delta T$ (i.e. $\Delta T$ induced by $\Delta I_D$) and the $\delta P$ (the additional power dissipated due to $\Delta I_D$) will simply yield the thermal resistance of the fin as follows:

$$R_{TH\_FIN} = \frac{\partial \Delta T}{\partial P} \ . \tag{6.3}$$

From Fig. 204 (b), it can be seen that the slope corresponds excellently with initial part of the $\Delta T$-$P$ from Fig. 202 (b), yielding us the $R_{TH\_FIN}$.

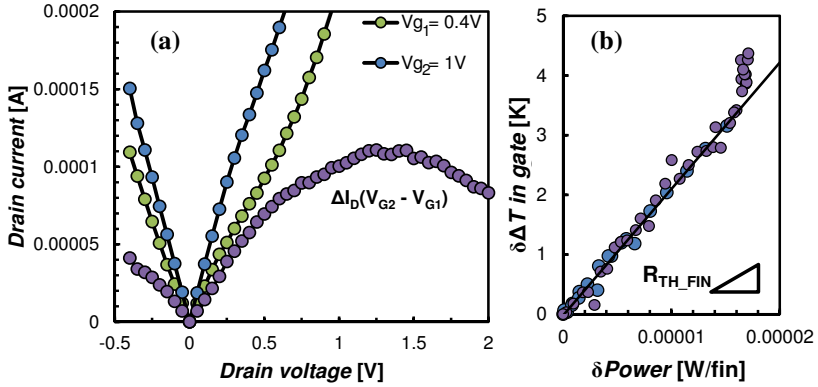



**Fig. 204: (a) The $\Delta I_D$ yields only channel modulated current without parasitic leakage current. (b) The slope of $\delta\Delta T$-$\delta P$ corresponds excellently with initial part of the $\Delta T$-$P$ from Fig. 202.**

In Fig. 205, we show the derivative of the original $\Delta T/dP$, which corresponds to the local $R_{TH}$. For both $V_G$ values, the data are converging to the same $R_{TH}$ at elevated $V_D$, which yields us the value of the thermal resistance of the parasitic channel ($R_{TH\_PARASITIC}$).

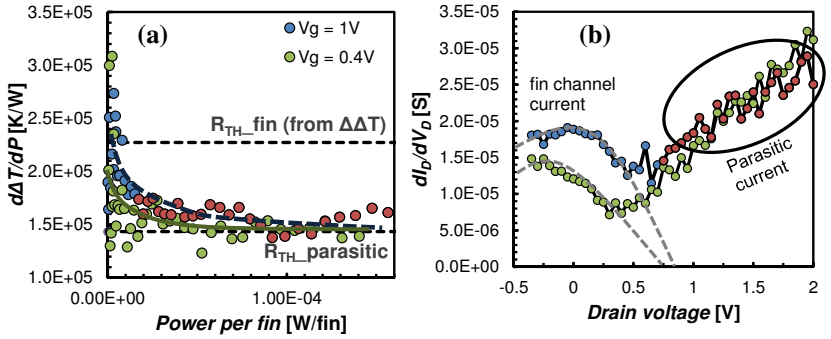



**Fig. 205: (a) $R_{TH\_PARASITIC}$ can be extracted from the saturation behavior of $d\Delta T/dp$ at high current levels. (b) The $dI_D/dV_D$ shows a strongly increased noise at elevated $V_D$, being the indication of a parasitic channel located at a defective region.**

Fig. 205 (b) shows the output conductance of the device, which shows a noisy behavior in the region where the parasitic current occurs (at high $V_D$). This is an indication that the parasitic current is located nearby a defective (i.e. with a lot of charge-traps) region in the device. This is corroborated by simulations in Fig. 206, where the thermal resistance of a secondary heat source in the device at various locations is compared to the measured $R_{TH\_PARASITIC}$. Locating the parasitic channel in the InGaAs close to the InP interface is the best approximation for the parasitic channel location.

We can thus conclude that the parasitic channel is due to defects, probably caused during the epitaxial growth of the InGaAs channel on the InP buffer, which is probably too thin to capture the lattice mismatch as SRB.

The results of benchmark, corrected for the parasitic fin and comparing the InGaAs/InP and InGaAs/GaAs FinFET are finally shown in Fig. 207 and discussed in the next Section.

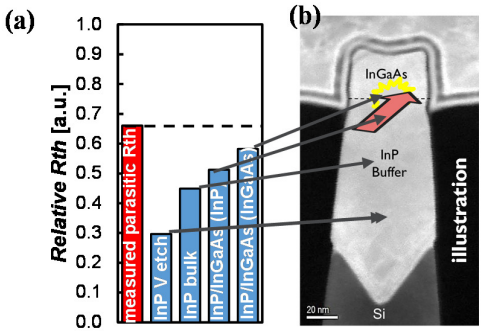



**Fig. 206: (a) Measured versus simulated thermal resistance of the parasitic channel relative to the 'real' channel $R_{TH}$. (b) TEM micrograph illustrating the various locations that are simulated. Locating the parasitic channel in the InGaAs close to the InP interface is the best approximation for the parasitic channel location.**

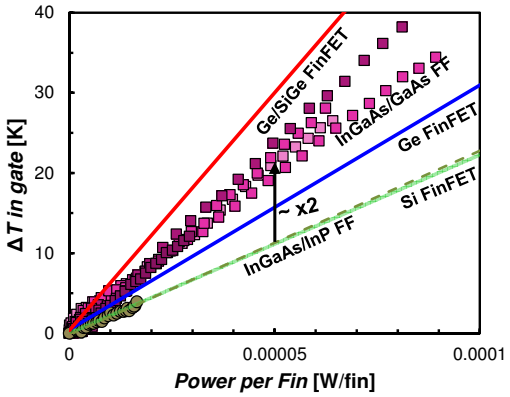



**Fig. 207: Comparison of the thermal resistances of the InGaAs/GaAs FinFET and the InGaAs/InP FinFET for $N_{FIN}$ = 10 and $W_{FIN}$ = 20nm. After correction for the parasitic channel, the InP buffer FinFET shows to have a $R_{TH}$ similar to the Si FinFET devices.**

### *6.5.3 Discussion*

The main observation from the measurement results in Fig. 207 is the ~x2 difference in thermal resistance between the GaAs buffer and InP buffer devices (after correction), whereas their bulk thermal conductivities are far less dissimilar than the SiGe versus Ge case.

This can be explained with the in-line TEM images in Fig. 208. It is clear that the buffer thickness for the InP devices is much smaller than the GaAs devices, i.e. only a small fraction of the Si fin is replaced by InP. For that reason, the InGaAs/InP devices show very similar behavior to full Si FF.

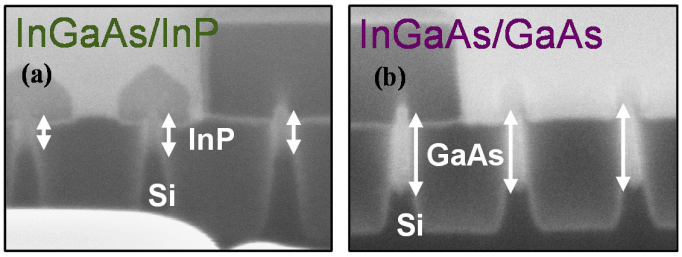



**Fig. 208: In-line TEM of (a) the InP buffer devices and (b) the GaAs buffer devices, showing a large difference in buffer ratio. The InP device is mostly a Si FinFET.**

A large spread on the data of the GaAs buffer is observed, which we attribute to the varying thickness of the GaAs buffer across the wafer, as is shown in the in-line TEM data in Fig. 209. The edge devices have a higher $R_{TH}$ than center devices. Counterintuitively, the effect on the thermal resistance is inversely proportional to the GaAs SRB thickness, since the actual *height* of the InGaAs channel is inversely proportional to the GaAs buffer height, i.e. the power density per unit footprint is strongly reduced for the devices with the thick buffer.

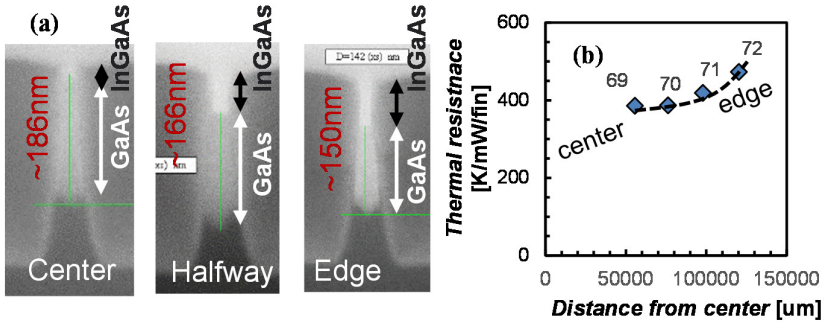



**Fig. 209: (a) In-line TEM showing the variation of GaAs buffer and InGaAs channel thickness across various devices on the wafer. (b) Edge devices have a higher thermal resistance than center devices.**

## 6.5.4 Conclusions

For III/V FinFET devices, we see a similar effect of changing the buffer material as in the SiGe/Ge SRB pFinFETs. The overly pronounced difference observed in the measurement data was attributed to non-equal thickness of the respective SRBs, and the presence of a parasitic channel in the InP devices.

## 6.6 III-V GAA-NW: impact of the nanowire

From the perspective of having a parasitic leakage path in previously described InP/InGaAs systems, the GAA-NW architecture is extremely attractive for the InP/InGaAs system. By selectively removing the InP buffer layer from underneath the InGaAs channel the main leakage path is eliminated and the electrostatic channel control of the NW could potentially be improved. In this section, we will refer to the number of *wires* and their widths *also* as $N_{FIN}$ and $W_{FIN}$, whereas in this case it consists of GAA-NW devices.

### *6.6.1 Device description*

The *first-generation* GAA-NW devices fabricated here follow the process flow of the III/V FinFETs as the basic structure [Waldron14]. Now, the InP is removed prior to the deposition of the high-k/metal gate stack to create GAA devices. Process details can be found in [Waldron15]. Alternatively, also devices with InAs channel were created, which can be of interest in this study as InAs does not suffer from alloy-scattering-reduced thermal conductivity [Palankovski00]. A schematic illustration of the cross-sections of the device at the channel and at the source/drain regions and a TEM at the gate region are shown in Fig. 210. This shows that in the ideal case, the InP is completely removed underneath the gate, and can be still present in the S/D region. An illustration along the fin is shown in Fig. 211.

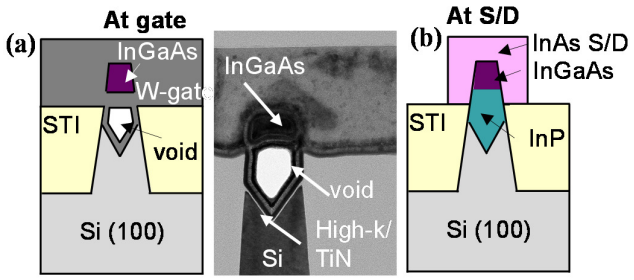



**Fig. 210: (a) Schematic and (b) TEM of the first generation III/V GAA-NW, which are generated by selectively etching the InP buffer in the channel region, creating a void. (c) Still InP is left in the trench in the S/D regions.**

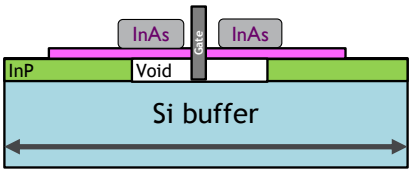



**Fig. 211: Apart from the region nearby the gate, InP is present in the rest of long (12.5um) fin trenches.**

Our *second generation* NWs follow a slightly different process. Prior to the deposition of the high-k stack the channel layer has been treated with a Wet HCl based Atomic Layer Etch (WHALE) digital etch process, mainly to remove any surface damage from previous processing. However, the WHALE process, which controllably removes ~1.4nm of InGaAs per cycle, was also found to be a good method to scale the nanowire diameter (Fig. 212). Moreover, these second generation nanowires show the absence of the void underneath the channel, which mean that in this case, the original top part of the fin is has been replaced with gate metal.

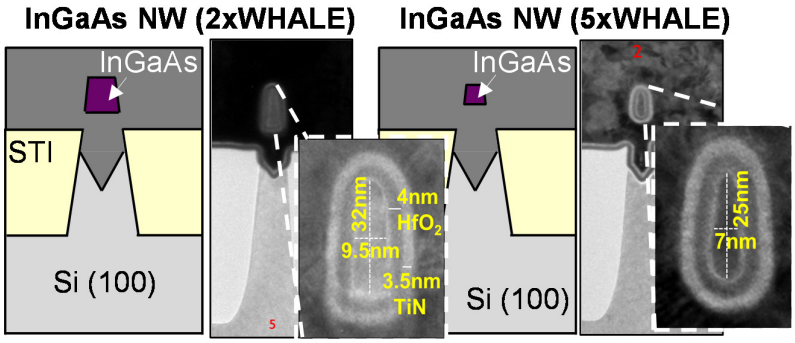



**Fig. 212: Schematic and TEM micrographs of the second generation GAA-NW, also illustrating the effect of the WHALE digital etch process on the average nanowire diameter. Note the absence of the void underneath the gate.**

### *6.6.2 Interpreting measurement data*

In a first set of measurements on the first generation GAA-NW, the thermal resistance appears to be correlated with $I_D$-$V_G$ data (Fig. 213). The devices with a *lower* off-state current $I_{OFF}$ (blue colored symbols) show an *increased* $R_{TH}$. The devices with a higher off-state current (green colored symbols) show a *decreased* thermal resistance.

This effect can be attributed to remaining InP in the trench under the gate. The remaining InP can give rise to a parasitic (non-gate-controlled) off current. Moreover, if InP is remaining in the trench, it is also a better thermal conductor than the void or the $SiO_2$ in the gate. The narrowest devices with $W_{FIN}$ of 20nm were not functional at all.

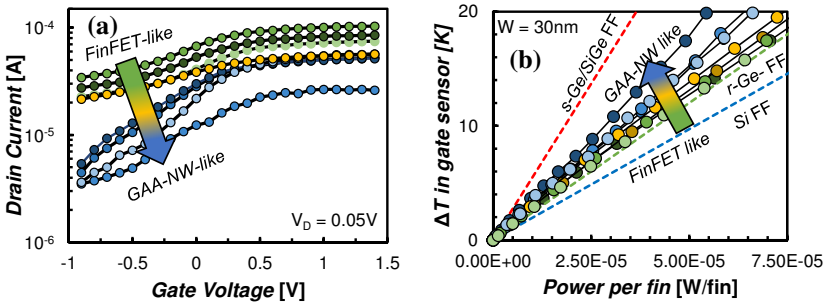



**Fig. 213: $I_D$-$V_G$ data of various nanowire devices of $W_{FIN}$ = 50nm, grouped into devices with high $I_{OFF}$ and lower $I_{OFF}$ due to parasitic current. (b) The thermal resistance of the devices appears to be correlated with the $I_{OFF}$.**

On average, the devices are shown to exhibit a $R_{TH}$ similar to the full Ge FinFET of Section 6.4. This can be explained by the fact that the devices are not *full* nanowires, i.e. surrounded with oxide or a void from the bottom over their entire trench length. If that was the case, $R_{TH}$ would be expected to be in the order of even worse than the SiGe devices (simulation results not shown here).

Measurements on the second generation of GAA-NW, with an optimized InP etch are then performed (Fig. 214). The narrowest devices with $W_{FIN}$=20nm and $N_{FIN}$=10 are now functional, albeit with a large variability. Surprisingly, the wide devices with multiple $N_{FIN}$ appear to have abnormal low thermal resistance. The reason of this is unknown, but can probably we attributed to other process related issues.

For systematic benchmarking of the NW process variations in the following paragraphs, we will therefore utilize only the narrowest devices with $N_{FIN}$ = 10 (since those are typically expected to have the highest sensitivity), unless otherwise noted.

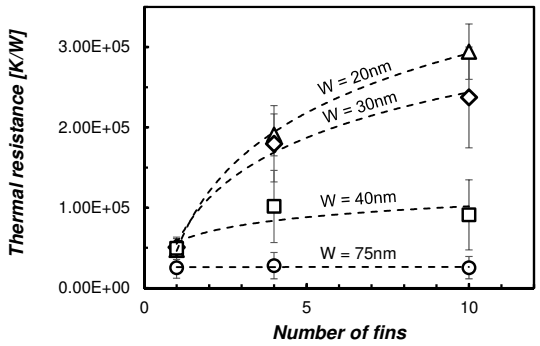



**Fig. 214: 20 and 30nm wide wires show expected $N_{FIN}$ and $W_{FIN}$ dependence whereas wide devices show no dependency on $N_{FIN}$.**

### *6.6.3 Measurement results*

Fig. 215 shows an illustration of the thermal network for a GAA-NW device and the measurement results obtained on our second generation NWs. Overall, the $R_{TH}$ of the nanowires is in the same order as typical full Ge FinFETs. We attribute this to the presence of the gate metal around (and underneath) the wire towards the fin, and the long fin trenches, where InP is still abundantly present at the contacts, illustrated in Fig. 211 earlier.

Subsequently, we investigate the effect of the nanowire scaling by WHALE etching, which reduces the nanowire diameter by 1.4nm per cycle. A 5x WHALE etching will therefore reduce the diameter by 7nm over the original diameter, which is a reduction of about 15%. Overall, it appears that scaling of the III/V channel has very little effect on the average extracted $R_{TH}$. A small *decrease* in $R_{TH}$ can be observed, albeit within error bar.

An explanation for this counterintuitive result is that the heat stays confined in scaled wires. By reducing the channel surface, while keeping the gate volume constant, i.e. thus increasing $R_{TH\_GATE}$ in Fig. 215(a), the observed temperature in the gate will be decreasing. This is confirmed by 3DFEM simulations in Fig. 216, which also show a decreased *observed* thermal resistance as the channel diameter is reduced in cycles of 1.4nm, whereas the temperature of the wire itself will *increase.*

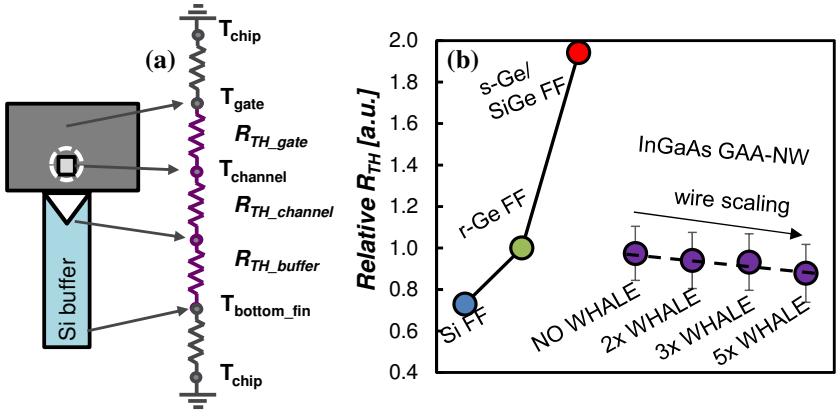



**Fig. 215: (a) Equivalent thermal network for the GAA-NW devices and (b) scaling the channel diameter with the *WHALE* etching process is *reducing* the observed thermal resistance.**

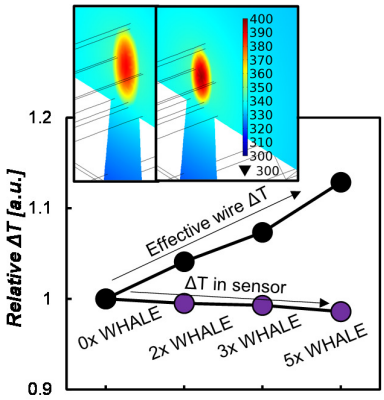



**Fig. 216: 3DFEM simulation result of the impact of channel scaling (-1.4nm per WHALE cycle) on the temperature of the gate and the temperature of this wire.**

A very similar effect is observed when comparing the InGaAs channel devices with InAs channel NW. TEM micrographs with Energy Dispersive X-Ray (EDX) data, show the both channel materials in Fig. 217, i.e. it shows

the InGaAs channel and but also the absence of Ga in the InAs channel FET cross-section.

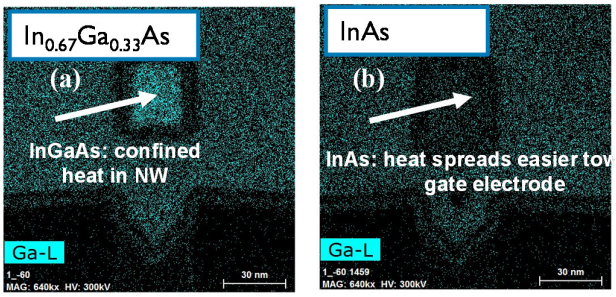



**Fig. 217: TEM micrograph combined with Energy Dispersive X-Ray (EDX) showing the (a) presence of Ga in the InGaAs channel FET and the (b) absence of it in the InAs FET.**

The measurement results in Fig. 218 show that the InAs channel—which has intrinsically a better thermal conductivity than the InGaAs alloy—tends to slightly but systematically *increase* the observed thermal resistance by 9% and 36% for moderately and maximally scaled wires respectively.

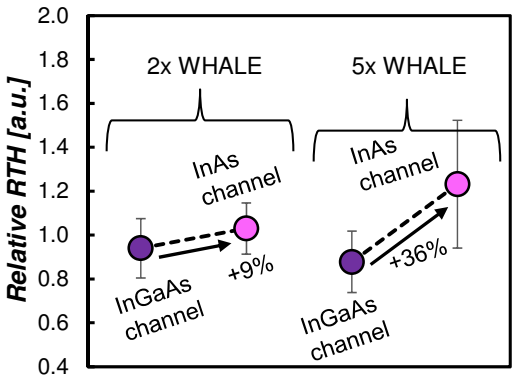



**Fig. 218: Thermal resistance for InAs channel FETs compared to InGaAs for moderately (2x WHALE) and maximally scaled (5x WHALE) nanowires, all showing an apparent increase in thermal resistance for InAs devices.**

Also these obtained results can be explained in a similar way as channel scaling: the good thermally conductive InAs channel has a better heat transfer of NW towards the gate electrode, thus a lower $R_{TH\_GATE}$.

It appears thus that the increase of the self-heating effect *within the nanowires*, either due to a reduced channel surface by wire scaling or due reduced thermal conductivity, can no longer be observed with gate resistance thermometry alone. Simulations are thus required if the effect on the thermal resistance of channel materials or channel diameters is to be assessed. If the devices show decent electrostatic behavior, also indirect measurement techniques (such as RF-measurements) can be utilized to assess these intrinsic heating effects.

### *6.6.4 Conclusions*

The overall thermal resistance of GAA-NW devices was shown to be comparable to Ge FinFET devices due to abundant presence of InP in long trenches. However, both scaling the NW channel by the WHALE digital etching process and reducing the thermal conductance of the channel by comparing InGaAs and InAs channels tends to reduce the *extracted thermal resistance of the gate*, which we attribute to the reducing heat transfer between the wire and the gate.

## 6.7 Conclusions

In this Chapter, our test-structures were proven accurate for self-heating assessment of novel fin and gate-stacks. Simulations and measurements showed that high-mobility materials and particularly alloys, such as a SiGe buffer for strain-induced mobility enhancement in Ge, can bring along a severe penalty on the thermal properties of the devices.

By making projections for the SiGe/Ge FinFETs, it was shown that the effect of mass disorder scattering is mitigated with scaling due to the more confined geometry of the fins, i.e. the impact of using materials with lower thermal conductivity will be smaller in nanoscale structures than in bulky devices.

It was shown that analyzing thermal properties combined with electrical measurements are powerful additional tools for analyzing intrinsic (e.g. InP or GaAs buffer thermal properties and thickness variations) and parasitic behavior (e.g. physically localizing leakage path due to EPI defectivity) of devices.

Considering nanowire devices, it was observed that the gate resistance thermometry reaches its limits in detecting the actual channel $\Delta T$ variations. Scaling the nanowires which should undoubtedly yield an elevated power density and higher temperature in the wire, resulted in the observation of *decreasing* thermal resistances, attributed to the reduced heat transfer between the wire and the gate.

## 6.8 References

[Bury15] Bury E., et al., "Characterization of Self-heating in High-Mobility Ge FinFET pMOS devices", in VLSI Symp. Tech. Dig., pp 160-161, (2015).

[Cheng02] Cheng Y-K., et al., "Electrothermal Analysis of VLSI Systems", Springer, 2002.

[Eneman12] Eneman G., et al., "Stress Simulations for Optimal Mobility Group IV p- and nMOS FinFETs for the 14 nm Node and Beyond", in Proc. IEDM, pp. 131-134, (2012).

[Fitzgerald91] Fitzgerald E.A. and Chand N. "Epitaxial necking in GaAs grown on pre-pattemed Si substrates", in Journal of Electronic Materials, Vol. 7, No. 7, pp. 839-853, (1991)

[Franco16] Franco J., et al, "Demonstration of an InGaAs gate stack with sufficient PBTI reliability by thermal budget optimization, nitridation, high-k material choice, and interface dipole", in IEEE VLSI Symp. Tech. Dig., pp. 42-43, (2016).

[Mautry90] P. G. Mautry and J. Trager, "Self-heating and temperature measurement in sub-pm-MOSFET's," in Proc. IEEE 1990 Int. Conf. Microelectronic Test Structures, vol. 3, pp. 221-226, (1990).

[Mitard14] Mitard J. et al, "15nm-WFIN High-Performance Low-Defectivity Strained-Germanium pFinFETs With Low Temperature STI-Last Process", in VLSI Tech. Dig., pp. pp 138-139, (2014).

[Palankovski00] Palankovski V., "Simulation of Heterojunction Bipolar Transistors", PhD Dissertation, Technische Universität Wien, available online: http://www.iue.tuwien.ac.at/phd/palankovski/ (2000).

[Waldron14] Waldron N., et al, "An InGaAs/InP Quantum Well FinFET Using the Replacement Fin Process Integrated in an RMG Flow on 300mm Si Substrates", in VLSI Symp. Tech. Dig., pp. 32-33, (2014).

[Waldron15] Waldron N. et al, "Gate-All-Around InGaAs Nanowire FETS with Peak Transconductance of 2200μS/μm at 50nm Lg using a Replacement Fin RMG Flow", in Proc. IEDM, pp. 799-802, (2015).

[Witters13] Witters L., et al., "Strained Germanium quantum well pMOS FinFETs fabricated on in situ phosphorus-doped SiGe strain relaxed buffer layers using a replacement Fin process", in Proc. IEDM, pp. 20.4.1.-20.4.4, (2013).

[Zhou16] Zhou D. et al., "Scalability of InGaAs Gate-All-Around FET integrated on 300mm Si platform: Demonstration of channel width down to 7nm and Lg down to 36nm", in VLSI Tech. Dig, pp. 166-167, (2016).

# Chapter 7: Conclusions and perspectives

In this Chapter, we will highlight the conclusions that could be made from the work performed in this thesis. Furthermore, we will provide interesting and challenging topics that remain to be investigated.

## 7.1 Conclusions

Below, we highlight the main conclusions of this thesis:

- Measurement techniques were developed and presented, which allowed the extraction of time-zero properties of UT-EOT devices: the single pulse *C*-*V*-technique was presented in which the *C*-*V* characteristics of leaky devices can be extracted, whereas a CBCM circuit was presented, capable of extracting device capacitances of nanoscale devices.

- Related to the long term-BTI we proposed and corroborated the *C*-*V*-eMSM technique, based on the conventional eMSM technique. This technique was found to be very useful in systematic BTI evaluation of novel high-k gate stacks, and is used to assess BTI in capacitors up to date.

- Using the *C*-*V*-eMSM technique and in the pursuit of improving the NBTI lifetime, we found that gate stack annealing conditions play an important role for gate stack quality. We have shown that the initial differences in BTI reliability between gate first and gate last processing approaches could be attributed to thermal conditions. We have shown that by applying the right annealing conditions for gate last stacks, they can get up to the reliability of their gate first counterparts.

- The fundamental origin of the accelerated BTI trend appears to be related to the scavenging of the $SiO_2$ interfacial layer, causing oxygen vacancies near the $SiO_2/HfO_2$ interface, which get charged by holes coming from the channel in the high-k. Given the experimental observations and the correlation with the EWF roll-off, it appears that these fixed charges in turn create a defect band-offset for the high-k, such that the high-k defect level becomes more accessible.

- Extended analysis of currents on all terminals of nanoscale devices yields large insight in TAT/SILC, RTN and BTI mechanisms. We proposed a model capable of explaining the measurement observations, and in particular the positive and negative $I_D$ and $I_G$ correlations. We found that a refined multi-state defect model, based on the model proposed by Grasser *et al.*, can explain correlated and non-correlated gate currents.

- A methodology was shown allowing the extraction of the physical position of the leakage paths in inversion in nanoscale devices, where the current-ratio technique starts to lose resolution.

- We found that in HKMG, most "BTI visible defects" do not have a large contribution to SILC current and vice versa, SILC defects do not show a large contribution to the $V_{TH}$ shift. Even though stress can result in defect generation, it will not necessarily result in activation of leakage paths, but it can also de-activate these paths.

- Measurement methodologies for measuring the self-heating effects were discussed and assessed. We proposed a new methodology based on a heater-sensor configuration and corroborated those with finite-element simulations in planar devices.

- We showed that enhanced electro-thermal Mont-Carlo simulation techniques require proper boundary conditions to yield correct results and proposed multi-scale simulations to solve this issue. We showed that the heater-sensor structure allows to sense the self-heating in the device and also the drain-hotspot by reversing the bias conditions and corroborated this with multi-scale simulations. We showed that the heater-sensor technique becomes less sensitive for more scaled technologies due to reduced heat transfer.

- We experimentally studied the impact of architectural changes on the thermal resistance of Si FinFETs. The EPI *composition* (e.g. SiGe S/D in PMOS) and the recess depth of the local interconnect around the S/D junction can have a significant impact on the $R_{TH}$. We showed that increasing the fin height by a deeper gate recess reduced the thermal resistance. The $R_{TH}$ of imec's first generation stacked GAA-NW is 60% elevated with respect to the baseline FinFET.

- We discussed a methodology to assess the impact of the self-heating effect on circuit performance by incorporating parameters in a BSIM4 model. We showed that the effect on circuit delay will depend on many operating conditions. Regarding device reliability, we showed that the impact of the self-heating effect on CHC is convoluted with other geometry-dependent effects, and thus cannot be straightforwardly extracted.

- We presented the gate resistance thermometry technique as an accurate technique for self-heating assessment of novel fin and gate-stacks, but also found that this technique reaches its limits in detecting the actual local channel $\Delta T$ variations in nanowires.

- We have shown by simulations and measurements that high-mobility materials and particularly alloys, such as a SiGe buffer for

strain-induced mobility enhancement in Ge, can bring along a severe penalty on the thermal properties of the devices.

- By making projections for the SiGe/Ge FinFETs, we have shown that the effect of mass disorder scattering will be mitigated with scaling due to the more confined geometry. The impact of using materials with lower thermal conductivity will thus be smaller in nanoscale structures than in bulky devices.

- We have shown that analyzing thermal properties combined with electrical measurements is a powerful additional tool for analyzing intrinsic (e.g. InP or GaAs buffer thermal properties and thickness variations) and parasitic behavior (e.g. physically localizing leakage path due to EPI defectivity) of devices.

## 7.2 Perspectives and future work

- The introduction of FinFET devices seems to have alleviated the BTI problem by improved electrostatic channel control (thus no longer necessitating an extremely scaled oxide) and a reduced oxide electric field because of the depletion of the oxide. For nanowire devices, the intrinsically reduced oxide field due to depletion might be counter-acted by a higher electric field due to their cylindrical shape. Optimizing gate stacks for BTI reliability will thus remain a crucial enabler for future technologies.

- In order to improve BTI reliability for gate stacks with an EOT below 1nm, more fundamental solutions will have to be found, most probably by no longer relying on oxygen scavenging to reduce the interfacial layer thickness. Example of such an alternative methodology is *the creation of an atomically flat silicon oxide monolayer*. For alternative channel materials such as Ge but mostly III/V devices*, our developed C-V-eMSM methodology can be used*

*to for efficient short-loop screening of gate stacks*. The most promising gate stacks can then be produced to full transistor lots.

- It was observed that gate resistance thermometry reaches its limits for assessing SHE in GAA-NW devices. in detecting the actual channel $\Delta T$ variations. Scaling the nanowires which will undoubtedly yield an elevated power density and higher temperature in the wire. This can only be observed by a technique which is sensitive to the channel $\Delta T$, such as s-parameter extraction by RF-measurements. In order to successfully complete such devices, *RF modules should be designed and the design should be such that it avoid parasitic gate resistance effects in the signal*.

- We proposed a methodology to verify the impact of self-heating effects on circuit performance by implementing parameters in a BSIM model. These *simulations could be corroborated with experimental data of GAA-NW transistors with or without high-mobility* channel materials.

- In order to assess the impact of self-heating effects on device reliability and reliability projections, we propose to use use arrays of devices in very high density, which allows parallel or separate stressing of the devices, capable of mimicking stress conditions in realistic low $N_{FIN}$ and the typically used high $N_{FIN}$ devices. The statistics of individual CHC degradation of many small devices stressed sequentially or in parallel should be obtained and compared. For this purpose, *we propose to design an array consisting of thousands of active DUTs which can be controlled separately*, but taking care of potential impacts of series resistances effects on common drain or source nodes.

- Recent trends in System-on-Chip (SoC) design point towards 3D-integration of multiple components. One of the options

comprehends the processing of logic devices throughout multiple layers in the wafer. *Logic devices placed in the middle- or back-end-of-line could severely be impacted by additional self-heating effects*, because they are surrounded by lowly conductive STI or back-end filler. This would form a very interesting case to assess the impact of self-heating effects on various methodologies of chip-stacking.

- Related to this, the thermal properties of the chip back-end materials and fillers should be studied, including the impact of placing *Through-Silicon-Vias (TSV) nearby devices. TSVs, which are typically made of Cu, could act as heatsink between parallel layers of the chips*, and could therefore help to mitigate the self-heating effect in stacked 3D-SoCs.

# List of Publications

## Publications as first author

## Awards

**IPFA 2014 Best Paper (Reliability) award**

**Bury E**., Degraeve R., Cho M., Kaczer B., Goes W., Grasser T., Horiguchi N. and Groeseneken G. "*Study of (correlated) Trap Sites in SILC, BTI and RTN in SiON and HKMG Devices*"

## Patents

**Bury E.**, Degraeve R., Hellings G., Franco J. and Kaczer. B "*Breakdown-based physical unclonable function*" – filed July 2016

## Invited International Conference talks

**ESREF 2015 – Berlin, Germany**

**Bury E.**, Degraeve R., Cho M., Kaczer B., Goes W., Grasser T., Horiguchi N. and Groeseneken G. "*Study of (correlated) Trap Sites in SILC, BTI and RTN in SiON and HKMG Devices*"

## International Conference proceedings

**Bury E.,** Kaczer B. Arimura H., Toledano Luque M, Ragnarsson L. Å., Roussel P., Veloso A., Chew S.A., Togo M., Schram T. and Groeseneken G., "*Reliability in Gate First and Gate Last Ultra-*

*Thin-EOT Gate Stacks Assessed with CV-eMSM BTI Characterization*", in Proc. International Reliability Physics Symposium, pp. GD.3.1-3.5, (2013).

**Bury E.**, Degraeve R., Cho M., Kaczer B., Goes W., Grasser T., Horiguchi N. and Groeseneken G. "*Study of (correlated) Trap Sites in SILC, BTI and RTN in SiON and HKMG Devices*", in Proc. IEEE IPFA, pp 250-253, (2014).

**Bury E.**, Kaczer B., Roussel P.J., Ritzenthaler R., Raleva K., Vasileska D. and Groeseneken G., "*Experimental validation of self-heating simulations and projections for transistors in deeply scaled nodes*" in Proc. IEEE International Reliability Physics Symposium, pp. XT.8.1 – XT.8.6, (2014).

**Bury E.**, Kaczer B., Mitard J., Collaert N., Khatami N.S., Aksamija Z., Vasileska D., Raleva K., Witters L., Hellings G., Linten D., Groeseneken G. and Thean A., "*Characterization of Self-Heating in High-Mobility Ge FinFET pMOS devices*", in Proceedings of the Symposium on VLSI Technology, pp. T60-61, (2015).

**Bury E.**, Kaczer B., Linten D., Witters. L, Mertens H., Waldron N., Zhou X., Collaert N., Horiguchi N., Spessot A. and Groeseneken G., "*Self-heating in FinFET and GAA-NW using Si, Ge and III/V channels*", **accepted** for International Electron Devices Meeting (IEDM) (2016).

# International Conference abstracts

**Bury E.**, Kaczer B. Arimura H., Toledano Luque M, Ragnarsson L. Å., Veloso A., Chew S.A., Togo M., Schram T. and Groeseneken G., "*Reliability in Gate First and Gate Last Ultra-Thin-EOT Gate*

*Stacks Assessed with CV-eMSM BTI Characterization*", presented at 43rd Semiconductor Interface Specialists Conference, December, (2012).

# Co-authored publications

## Book chapters

Cho M., **Bury E.**, Kaczer B. and Groeseneken G., "*Channel Hot Carrier degradation and Self-Heating Effects in FinFETs*" in Hot Carrier Degradation in Semiconductor Devices, Springer, 2015, pp. 287-308.

## International Journal Papers

Kaczer B., Franco J., Weckx P., Roussel P.J., **Bury E.**, Cho. M., Degraeve R., Linten D., Groeseneken G., Kukner H., Raghavan P., Catthoor F., Rzepa G., Goes W. and Grasser T., "*The defect-centric perspective of device and circuit reliability – From individual defects to circuits*", accepted for Solid-State-Electronics.

## International Conference Proceedings

Jang D., **Bury E.,** Ritzenthaler R., Garcia Bardon M., Chiarella T., Miyaguchi K., Raghavan P., Mocuta A., Groeseneken G., Mercha A., Verkest D. and Thean A. "*Self-heating on bulk FinFET from 14nm down to 7nm node*" in IEEE International Electron Devices Meeting (IEDM), pp. 11.6.1-11.6.4, (2015).

Lin. D., Alian A., Gupta S., Yang B., **Bury E.**, Sioncke S., Degraeve R., Toledano Luque M., Krom R., Favia P., Bender H., Caymax M., Saraswat K.C., Collaert N. and Thean A., "*Beyond interface: The*

*impact of oxide border traps on InGaAs and Ge n-MOSFETs*" in Proc. IEEE Electron Devices Meeting (IEDM), pp. 28.3.1 – 28.3.4, (2012).

Franco J., Kaczer B., Roussel P.J., **Bury E.**, Mertens H., Ritzenthaler R., Grasser T., Horiguchi N., Thean A. and Groeseneken G., "*NBTI in Si0.55Ge0.45 cladding p-FinFETs: Porting the superior reliability from planar to 3D architectures*" in Proc. IEEE International Reliability Physics Symposium, pp. 2F.4.1-2F.4.5, (2015).

Weckx P., Kaczer B., Chen C., Franco J., **Bury E.**, Chanda K., Watt J., Roussel P.J., Catthoor F., Groeseneken G., "*Characterization of time-dependent variability using 32k transistor arrays in an advanced HK/MG technology*", in IEEE Proc. IRPS, pp 3B.1.1-6, (2015).

Raleva K., **Bury E.**, Kaczer B. and Vasileska D., "*Uncovering the temperature of the hotspot in nanoscale devices*", in 2014 International Workshop on Computational Electronics (IWCE), pp. 1-3, (2014)

Putcha V., **Bury E.**, Weckx P., Franco J., Kaczer B., Groeseneken G. "*Design and simulation of on-chip circuits for parallel characterization of ultrascaled transistors for BTI reliability*", in IEEE Proc. International Integrated Reliability Workshop (IIRW), pp. 99-102, (2014).

Ritzenthaler R., Schram T., **Bury E.**, Mitard J., Ragnarsson L. -Å., Groeseneken G., Horiguchi N., Thean A., Spessot A., Caillat C., Srividya V. and Fazan P., "*Low-power DRAM-compatible Replacement Gate High-k/Metal Gate stacks*" in Proceedings of the European Solid-State Device Research Conference (ESSDERC), pp. 242-245, (2012).

Weckx P., Kaczer B., Franco J., Roussel P.J., **Bury E.**, Subirats A., Groeseneken G., Catthoor F., Linten D., Raghavan P., and Thean A., "*Defect-centric perspective of combined BTI and RTN time-dependent variability*", in IEEE Proc. International Integrated Reliability Workshop (IIRW), (2015).

Kaczer B., Franco J., Weckx P., Roussel P.J., **Bury E.**, Cho. M., Degraeve R., Linten D., Groeseneken G., Kukner H., Raghavan P., Catthoor F., Rzepa G., Goes W. and Grasser T., "*The defect-centric perspective of device and circuit reliability – From individual defects to circuits*", in IEEE Proc. ESSDERC, pp. 218-225, (2015).

Arimura H., Ragnarsson L. -Å., Schram T., Albert J.; Kaczer B., Degraeve R., **Bury E.**, Aoulaiche M., Kauerauf T., Thean A., Horiguchi N, Groeseneken G., "*Guidelines for reducing NBTI based on its correlation with effective work function studied by CV-BTI on high-k first MOS capacitors with slant-etched SiO2*", in Proc. IEEE International Reliability Physics Symposium, pp. 3C.4.1-3C.4.6, (2014).

Qazi S. S., Shaik A. R., Daugherty R. L., Laturia A., Vasileska D., Guo X., **Bury E.**, Kaczer B. and Raleva K. "*Multi-scale modeling of self-heating effects in silicon nanoscale devices*" in IEEE 15th International Conference on Nanotechnology (IEEE-NANO), pp. 1461-1464, (2015).

Kaczer B., Franco J., Cho M., Grasser T. Roussel P.J., Tyaginov S., Bina M., Wimmer Y., Procel L.M., Trojman L., Crupi F., Pitner G., Putcha V., Weckx P., **Bury E.,** Ji Z., De Keersgieter A., Chiarella T., Horiguchi N., Groeseneken G. and Thean A., "*Origins and implications of increased channel hot carrier variability in nFinFETs*", in IEEE Proc. IRPS, pp. 3B.5.1-6, (2015).